이학박사 학위논문

유전자 수준 변이 부담을 적용한 약물유전자와 질병유전자의 특성 분석

Genomic Characterization of Pharmacogenomic and Disease Genes using Gene-wise Variant Burden: evidence of utility in the field of computational pharmacogenomics

2020 년 08 월

서울대학교 대학원

협동과정 생물정보학과 생물정보학전공

박 유 미

A thesis of the Degree of Doctor of Philosophy

Genomic Characterization of Pharmacogenomic and Disease Genes using Gene-wise Variant Burden: Evidence of Utility in the Field of Computational Pharmacogenomics

유전자 수준 변이 부담을 적용한 약물유전자와 질병유전자의 특성 분석

August 2020

The Department of Bioinformatics,

Seoul National University

Yoomi Park

Genomic Characterization of Pharmacogenomic and Disease Genes using Gene-wise Variant Burden: Evidence of Utility in the Field of Computational Pharmacogenomics

by

Yoomi Park

A thesis submitted to the Department of Bioinformatics
in partial fulfillment of the requirements for
the Degree of Doctor of Philosophy in Bioinformatics at
Seoul National University

August 2020

Approved by Thesis Committee:

Professor Taesung Park Chairman

Professor Ju Han Kim Vice chairman

Professor Buhm Han

Professor Murim Choi

Professor Ji Wan Park

초 록

차세대 시퀀싱 기술이 발전함에 따라 전장 유전체에서의 변이 분포를 확인하는 것이 가능해졌다. 전통적인 단일 변이 기반 분석법은 드물게 발견되는 변이에 대한 통계적 파워가 떨어질 뿐만 아니라, 후보 변이가 발견되었다 하더라도 발견되는 모든 변이에 대해 그 기능적 중요도를 체계적으로 평가하기 어렵다는 점에서 변이와 표현형 간 연관 관계를 탐색하는데 한계가 있었다. 이러한 문제를 해결하기 위한 방법 중 하나로 특정 유전자 (혹은 영역)에서 발견된 변이들의 누적 효과를 통합하여 유전자 수준에서 표현형과의 연관 관계를 탐색하는 다양한 접근법이 제안되었다. 그 중, 통계 검정과 인구집단 수준의 유전자 점수는 접근 방법에는 차이가 있지만 유전자와 표현형 간 연관 관계를 탐색한다는 점에서 공통점이 있다. 반면, 유전자 수준의 변이 부담 점수 (GVB, Gene-wise Variant Burden score)는 주로 약물 유전체 분야에서 약물 부작용과의 연관성이 잘 정립된 유전자에서의 환자 개인에 대한 위험도를 예측하는 분석에서 다양하게 활용되었는데, 아직까지 그 유용성에 대해서 체계적으로 평가된 적이 없었다. 본 연구는 약물

유전체 분야 뿐만 아니라 희귀 질환, 복합 질환에서 GVB 의 평가적, 탐색적 유용성을 평가하는 것을 목적으로 한다.

이를 위하여, 1000 지놈 프로젝트로 부터 얻은 2,504 명의 건강한 사람의 전장 유전체 시퀀싱 데이터와 320 명의 소아 백혈병 환자에 대한 엑솜 시퀀싱 데이터가 사용되었다. 모든 단백질 coding 유전자에 대하여, 각 유전자 내에서 발견된 변이의 위해도를 *in silico* 예측 점수를 통해 평가한 뒤, 유해하다고 판단되는 모든 변이의 효과를 통합하는 유전자 수준 변이 부담 점수를 계산하였다. 소아 급성 림프모구성 백혈병 환자의 6-mercaptopurine (6-MP) 약물 독성 예측에서 유전자 수준 변이 부담 점수의 임상성 유용성을 평가하기 위하여, 소아 백혈병 환자에서 6-MP 약물에 대한 예측 대비 실제 투약 비율 (DIP, dose intensity percentage)을 수집하였다. DIP 를 임상적 종점 (clinical endpoint)으로 보고, 수신자 판단 특성 곡선 (Receiver operating characteristic (ROC) curve) 분석을 통해 유전자 수준 변이 부담 점수가 개인에 대한 약물 독성을 얼마나 잘 예측하는지 평가하였다. 민감도, 특이도, 양성 예측도, 음성 예측도, 그리고 정확도를 계산함으로써, 약물 유전체 분야에서 개인 약물 독성 예측에

사용되어온 가장 고전적인 방법인 스타 대립 유전자 기반의 하플로타이핑 방법론 (star allele-based haplotyping method)과 진단 정확도를 비교하였다. 더불어, 약물 부작용 이외에 희귀질환, 복합질환 등 다양한 유전 배경에서 유전자 수준 변이 부담 점수의 유용성을 평가하기 위하여, PharmGKB (The Pharmacogenomics Knowledge Base)와 DrugBank 데이터베이스로 부터 추출한 약물 유전자와 GAD (Genetic Association Database)로 부터 추출한 복합 질환 유전자, 그리고 OMIM (Online Mendelian Inheritance in Man)으로 부터 추출한 희귀질환 유전자를 사용하여 약물유전체 및 질병유전체에서의 유전자 수준 변이 부담 점수의 예측 성능을 평가하였다. 각 유전적 배경 (약물, 희귀, 복합 질환)에서 일곱개의 유전적 특성 (paralog 와 singleton 의 갯수, per-person mutability, CDS (coding sequence) 길이, PPI (protein-protein interaction) degree, 단백질 복잡도, 그리고 선택적 압력)의 분포 차이를 확인하고, 해당 특성을 반영하여 각 유전적 배경에 최적화 시킨 보정된 유전자 수준의 변이 부담 점수를 제안하였다. 추가로 320 명의 소아 백혈병 환자에 대한 엑솜 시퀀싱 데이터를 이용하여 *NUDT15* 과 *TPMT* 이외에 치오퓨린 약물 독성과 연관을 보이는 새로운 유전자

마커를 탐색하고, GVB 를 이용하여 새로운 유전자 마커의 단독 효과와 추가 효과를 평가하였다.

소아 백혈병 환자에서 유전자 수준의 변이 부담 점수를 활용한 6-MP 약물에 대한 독성 예측은 기존의 스타 대립 유전자 기반 분자 표현형 방법론과 비슷하거나 더 나은 결과를 보였다 (DIP ≤25 기준 AUC_{GVB}=0.677, $AUC_{star-allele}$ = 0.645). 유전자 수준 변이 부담 점수의 확장 가능성을 확인하기 위하여 다양한 유전 배경에서 그 성능을 평가해보면, 해당 점수는 약물 유전자 예측에 가장 효과적이었던 반면, 집단 수준의 점수를 제공하는 기존의 유전자 수준 방법론 (GDI, RVIS, 그리고 pLI)들은 희귀 질환 유전자 예측에 가장 효과적이어서 서로 상호 보완적인 역할을 한다는 것을 확인할 수 있었다. 특히 약물 유전자는 높은 paralog 갯수, 가장 낮은 PPI degree 와 짧은 CDS 길이를 보였던 반면, 복합질환 유전자는 가장 적은 singleton 갯수, 높은 양성 선택과 per-person mutability, 희귀질환 유전자는 낮은 paralog 갯수와 가장 긴 CDS 길이, 높은 선택적 압력과 낮은 per-person mutability 등, 유전적 배경 별로 매우 다른 분자 특성을 보였다. 특징적인 유전적 요소들의 효과를 반영했을 때, 유전자 수준 변이 부담

점수는 증가된 예측 성능을 보였다. 유전자 수준의 점수를 사용하여 새로운 바이오 마커인 *CRIM1* 의 치오퓨린 독성 예측 성능을 평가한 결과, 기존에 잘 정립된 *NUDT15* 과 *TPMT* 에 추가적인 영향은 물론이고 독립적인 영향도 보이는 것으로 확인되었다.

결론적으로, 서로 다른 유전적 배경을 가지는 표현형에는 각 특성을 반영하는 서로 다른 접근법이 필요하며, 유전자 수준 변이 부담 점수는 특히 집단 수준의 점수가 아니라 개개인에 대한 점수를 따로 제공한다는 이점이 있기 때문에 약물에 대한 반응성 차이가 사람 간 유전적 다양성으로 설명되는 약물 유전체 분야에서 그 쓰임이 가장 높을 것으로 판단된다.

주요어: 유전자 점수, 약물 유전자, 희귀 질환 유전자, 복합 질환 유전자, 변이 부담

학 번: 2014-21328

목 차

그림 목록

표 목록

1 서론

차세대 시퀀싱 기술이 빠르게 발달하면서, 임상적 중요도가 높은 유전 변이를 예측하고 목록화하는 것이 가능해졌다 [1-3]. 전장 유전체 연관 분석 (GWAS, Genome-wide association test) 에서 사용하던 전통적인 Fisher's exact test, Cochran armitage trend test 등의 방법은 단일 변이를 기반으로 유전형과 표현형 간 연관 관계를 확인하는 데 이용되어져 왔다. 전장 유전체 연관 분석에서 초점을 두었던 마커 변이 (tag SNP)의 경우, 특정 표현형에 원인이 되는 (causal) 변이보다는 해당 변이와 연관 (linkage disequilibrium) 되어 있는, 흔하게 발견되지만 직접적이지 않은 (indirect) 연관 관계를 갖는 변이를 탐색하는 것이 우선적인 목적이었기 때문에 단일 변이 기반 연관 분석 방법을 통한 탐색이 가능하였다 [4]. 그러나, 차세대 시퀀싱 기술을 통해 전장 유전체에 대한 유전적 조성을 확인하는 것이 가능해지면서 특정 표현형과 직접적인 (direct) 연관이 있는 변이를 탐색하는 데 중점을 두기 시작했다. 특히 인구집단 내 드물게 발견되는 변이는 그 기능적 중대성이 상대적으로 높고 [5], 희귀 질환은 물론이고 복합 질환에서도 두드러진 연관성을 보여 드문 변이와 표현형 간 연관 관계를 확인하기 위한 노력들이 시도되었으나, 통계적 파워가 떨어진다는 한계 때문에 샘플 수가 매우 크거나, 변이 효과가

매우 크거나, 혹은 변이 빈도가 너무 낮지 않은 이상 단일 변이 기반의 연관 분석으로는 변이와 표현형 간 연관성을 탐지하기 어려웠다 [6, 7].

최근 이러한 문제를 해결하기 위해서 유전자, 혹은 특정 영역 내 중요도가 높은 여러 변이의 효과를 통합하여 표현형과의 연관성을 탐색하는 방법론들이 제안되었다. 유전자 – 표현형 간 연관성을 탐색하기 위한 방법론은 크게 통계 테스트와 점수 기반 시스템의 두 카테고리로 분류할 수 있다.

1.1 통계 테스트

통계 테스트는 기본적으로 환자-대조군 디자인에 기반하여 그룹 간 빈도 차이를 보이는 특정 영역, 혹은 해당 영역에서 발견된 변이의 세트를 탐색하는 방법론이다. 특히 단일 변이 기반의 연관 검정으로는 탐지하기 어려운 드문 변이들의 효과를 통합하는 다양한 종류의 드문 변이 연관 분석법 (RVAT, Rare Variant Association Test)이 제안되었다 (표 1) [8]. 드문 변이 연관 분석법은 크게 검정에 포함되는 모든 변이가 표현형에 같은 방향 (위험을 높이거나 낮추는)으로 작용한다는 가정 하에 분석을 수행하는 단방향 회귀, 서로 다른 방향으로 작용하는 변이들이 섞여있다고 가정하는 양방향 분산 성분 테스트, 그리고 이 두 가지 방법을 혼합하여 가장 최적화된 선형 조합 조건을 찾는 혼합형

모델의 세 가지 카테고리로 분류할 수 있다. 단방향 회귀 유전자 기반 테스트의 경우, 특정 유전자 혹은 영역 내에서 발견된 변이들의 빈도를 단순히 통합 (collapsing) 하거나, 여기에 빈도의 역 (inverse)이나 변이의 관찰된 효과 (effect size)를 가중하여 통합하는 방법을 통하여 표현형과 유전 변이 세트간의 연관관계를 탐지한다. 분산 성분 테스트는 양방향 효과를 고려하는 방법론으로, 역시 대립 유전자의 빈도 등으로 가중치를 부여할 수 있다. 해당 두 가지 분석법을 적절히 조합한 최적화된 시퀀스 커널 연관 분석법 (SKAT-O, Optimized sequence kernel association test)은 현재 유전자 – 표현형 간 연관 관계를 탐지하는 데 가장 많이 활용되고 있다.

이러한 방법론들은 각각 검정에 포함되는 변이와 표현형 간 관계에 대한 기본 가정에는 차이가 있지만, 환자-대조군 디자인에 기반한 그룹 간 변이 부담 (variant burden) 차이, 혹은 연속 변수 검정의 최종 결과로서 변이 세트 수준의 유의 확률 (p 값)을 제공하게 된다.

표 1. 드문 변이 연관 분석법의 종류와 특성.

방법론 종류	방법론	방법론 이름	설명
단방향 희귀 변이 유전자 기반 테스트	결합 방법	다변량 및 축소 결합 (CMC)	모든 희귀 변이를 단일 변이 단위로 축소 결합함. 축소 결합된 변이의 개별 보유량 (dosage)을 표현형에 대해 회귀.
	가중 및 비가중 합계	변수 임계값 (VT)	환자 및 대조군에서의 드문 대립 유전자 수의 합. 포함 대상이 되는 변이의 대립 유전자 빈도 임계값은 테스트 통계량을 최대화하도록 조정됨.
		가중 합 방법 (WILCOX-WSS)	표현형과 **빈도의 역 (inverse)을 가중**한 드문 변이 점수 간의 Wilcoxon Rank Sum 테스트.
		커널 기반 적응 클러스터 (KBAC)	변이에 대한 **가중치가 관찰된 효과 크기에 기반하여 조정**됨. 가중 대립 유전자수의 합으로 점수를 매김.
	환자:대조군 총수 요약 방법	변이 부담 방법 (Burden)	환자 및 대조군에서의 순수 대립 유전자의 수 총 합을 비교한 순열 기반 검정.
양방향 분산 성분 유전자 기반 테스트	분산 성분 테스트	C-ALPHA	환자 및 대조군에서의 예측 대비 실제 발견된 변이 갯수의 편차를 탐지.
		시퀀스 커널 연관 분석 (SKAT)	**대립유전자 빈도로 가중치**를 부여한 일반화된 형태의 C-ALPHA 테스트
단방향 및 분산 성분 테스트의 선형 조합	단방향 및 분산 성분 테스트의 선형 조합	최적화된 SKAT (SKAT-O)	단방향의 변이 부담 검정과 분산 성분 SKAT 테스트의 적응성 선형 조합.

1.2 유전자 수준의 점수 기반 시스템

최근 제안된 점수 기반 시스템은 유전자 – 표현형 간 관계를 추정하기 위한 가장 강력한 접근 방법 중 하나로, 특정 목적에 맞는 (유전자, 약물, 혹은 pathway) 혹은 특정 영역 (유전자, 엑손, 혹은 도메인) 내에 존재하는 변이의 효과를 통합하여 수치적으로 제공하기 때문에 보다 직관적인 해석을 가능하게 했다 [9–12]. 유전자 수준 점수 기반 시스템은 크게 인구 집단 기반의 점수와 개인화된 점수로 나눌 수 있는데, 두 시스템은 활용 목적 면에서 매우 큰 차이를 가지고 있다.

1.2.1 인구 집단 기반의 점수 시스템

인구 집단 기반의 점수 시스템은 유전자 – 표현형 간 연관관계를 탐색하는 목적으로 사용된다는 점에서 기존의 통계 테스트들과 공통점이 있다. 초기 유전자 수준 점수 기반 시스템으로는 인구 집단 내 변이 빈도 분포를 바탕으로 계산되는 RVIS (Residual Variation Intolerance Score) [9], pLI (probability that a given gene falls into the Haploinsufficient category) [13], GDI (Gene Damage Index) [10] 등이 있다. RVIS 는 전체 유전자를 대상으로 인구 집단 내에서 예측 대비 실제 발견되는 기능 변이의 비율을 바탕으로 유전적 감내성 (genic intolerance)에 대한 순위를 매겨 수치적으로 환산한 점수이며, GDI 는

유전자 별 인구집단 내 변이 부담을 직접적으로 계산한 점수이다. pLI 와 RVIS 는 모두 유전적 감내성의 개념에 기반하는 점수이지만, RVIS 가 유전자 내 모든 종류의 기능 변이를 대상으로 하는 반면, pLI 는 영향력이 큰 기능 상실 변이 (loss of function variant) 의 분포만을 고려한다는 점에서 차이가 있다. 이러한 인구 집단 기반의 유전자 점수는 특정한 표현형이 없는 일반 인구집단에서 발견된 변이의 분포를 바탕으로 유전자를 특성화하여 점수를 산출한다. 자세하게는, 각 유전자에서 예측 대비 실제 발견되는 기능 유전 변이의 분포 차이 (잔차 값)를 유전자 점수로 할당하고, 이 때 유전자 점수 값이 작을수록 (예측 대비 실제 발견되는 기능 유전 변이의 수가 현저히 적은 경우) 해당 유전자는 유전 변이에 불내성 (intolerance)을 가지고, 반대로 큰 유전자 점수 값이 클수록 (예측 대비 실제 발견되는 기능 유전 변이의 수가 현저히 많은 경우) 유전 변이에 내성 (tolerance)을 가진다고 해석한다. 이 때 '불내성'의 특성을 가지는 유전자는 주로 발달 질환과 연관성이 많이 보고되는, 보존된 (conserved) 영역에 위치한 필수 (essential) 유전자일 확률이 높은 반면, '내성' 유전자의 경우 면역 질환과의 연관성이 보고된 유전자들과 유사한 특성을 가지는 것 확인할 수 있었다. 해당 방법론들은 주로 희귀질환 분석에 초점을 두고 활용되어져왔다. 이러한 인구 집단 기반의 점수 시스템은 일반 인구 집단에서 발견되는

변이의 분포에서 유전자를 특성화하고 그와 비슷한 특성을 가지는 새로운 유전자와 표현형 간의 연관성을 확인한다는 점에서 환자-대조군 디자인에 기반한 기존 통계 테스트와는 매우 상이한 접근 방법을 가진다.

1.2.2 개인화된 점수 시스템

인구 집단 기반의 유전자 점수는 개인 별 유전적 변동성을 체계적으로 고려하지 않는다는 한계가 존재하였다. 맞춤 의료의 실현을 목전에 두고 개인에 대한 유전적 차이를 체계적으로 비교하는 방법론에 대한 필요성이 높아졌는데, 특히 약물 유전체 분야에서는 특정 약물에 대한 사람 간 반응성의 차이가 유전 변이의 다양성으로 설명되기 때문에 개인에 초점을 두는 접근 방법의 개발이 요구되어 왔다 [14]. 개인화된 점수 시스템은 이미 특정 표현형과의 연관관계가 잘 정립된 유전자에서 발견된 변이 효과를 효과적으로 통합하여 특정 질병 혹은 표현형에 대한 개인의 위험도를 평가하는 것을 목적으로 한다는 점에서 유전자 – 표현형 간 연관 관계 탐색에 초점을 두었던 통계 테스트나 인구집단 수준의 점수와는 그 쓰임에 차이가 있다. 개인화된 점수 시스템의 가장 기본적인 형태는 유전 변이 부담 (genetic burden)으로, 1) 특정 영역 내에서 한 번 이상 기능 변이가 발견되는 사람과 그렇지 않은 사람을 나누는 이분법적 유전 변이 부담 계산법과 2) 특정 영역 내 발견된 기능

변이의 수를 합하는 방법, 그리고 3) 여기에 변이 별 효과 등의 가중치를 부여한 가중합 방법 등이 존재한다. 현재 가장 많이 사용되고 있는 개인화된 점수 시스템 중 하나는 다형성 위험 점수 (PRS, polygenic risk score)로 [15], 특히 복합 질환에서 개인의 위험도를 예측할 때 그 활용도가 높음이 확인되었다. 해당 점수는 전장 유전체 연관 분석의 결과인 요약 통계 (summary statistics)를 바탕으로 위험 대립 유전자 및 그 가중치를 정의하여 개인이 가지고 있는 위험 대립 유전자 수의 가중합을 계산하는 방법론이다 (표 2). 이러한 방법론은 기존의 연구를 바탕으로 질병 혹은 특성 별로 차별화된 점수를 제공할 수 있지만, 기존에 수행된 전장 유전체 연관 분석 결과의 신뢰도와 연구된 인종에 매우 의존적이라는 한계가 존재하였다.

최근 제안된 유전자 수준 변이 부담 (GVB, Gene-wise Variant Burden) 점수는 유전자 내 발견된 모든 기능 변이의 효과를 통합한 개인 별 유전자 점수로, 약물 유전체 분야에서 활발하게 사용되어져 왔다 [14, 16-19]. 차세대 시퀀싱 기술에 적합한 방법론으로 흔한 변이는 물론이고 드물거나 개인에서 새롭게 발견된 변이의 효과까지 통합하는 것을 목적으로 하며, 특히 진화적 압력에 따른 위해도를 바탕으로 가변성이 높은 영역 (variable region)에서 발견된 위험 변이 효과의 누적을 계산한다는 점에서 차별점을 갖는다. 유전자 수준 변이 부담

점수는 인구 집단이나 기존 연구에 독립적인 점수이지만, 서로 다른 질병이나 특성 간 변이 효과에 차별화를 두지 않는다는 점에서 한계가 있다.

표 2. 유전자 수준 변이 부담 점수와 다형성 위험 점수의 특성 비교.

	유전자 수준 변이 부담 점수 (GVB)	다형성 위험 점수 (PRS)
수식	i 번째 사람에 대한 조정되지 않은 유전자 수준 변이 부담 점수 $$GVB(G_i) = \begin{cases} 1 & , if\ n(G_i) = 0 \\ \left(\prod_{j=1}^{n} v_j\right)^{\frac{1}{n}} & , if\ n(G_i) > 0 \end{cases}$$ G_i: i 번째 사람의 유전자 v_j: j 번째 유전 변이에 대한 *in silico* 예측 변이 점수 $n(G_i)$: i 번째 사람에서 유전자 내 발견된 변이의 갯수	i 번째 사람에 대한 조정되지 않은 표준 다형성 점수 $$S_i = \sum_{j=1}^{M} X_{ji}\hat{\beta}_j$$ X_{ji} : i 번째 사람, j 번째 유전 변이에 대한 유전형 β_j : 개별 마커 변이에 대한 추정 효과
정의	유전자의 모든 코딩 변이체에 대한 *in silico* 점수 (예: SIFT 점수)의 기하 평균을 계산하여 집계 된 유전자 단위 효과	개인이 가지고있는 위험 대립 유전자 수의 가중 합, 위험 대립 유전자 및 그 가중치가 변이 및 전장 유전체 연관 분석에 의해 검출된 측정 효과에 의해 정의됨
점수 특성	유전자 기반, (사람 수 x 유전자) 개 점수 도출	영역 기반, (사람 수 x 표현형) 개 점수 도출
활용 분야	약물, 복합질환	복합질환
데이터 소스	차세대 시퀀싱 데이터 (WXS, WGS)	Chip 데이터 (GWAS)
개인화된 점수?	Y	Y
주요 변인	진화적 압력에 기반한 위해도	GWAS 요약 통계에서 추출한 변이 효과

한계	- 질병 혹은 특성 별 점수 차별화 없음 (서로 다른 표현형 간 기능적 영향을 구분하지 않음 [예: 병인성과 질병을 유발하지 않는 약물 연관 기능성 변이]) - 체계적 점수 보정 전략이 필요	- 기존 GWAS 분석의 신뢰도에 의존적 - 샘플 사이즈 의존적 (변이 기능 효과 결정) - 인종 의존적 (특정 인종에 대한 기능 효과에 대한 결과가 없다면 점수 계산 불가능)

1.3 유전자 수준 변이 부담 점수의 최적화

유전자 수준 변이 부담 점수의 활용을 위해서, 1) 계산에 포함 할 변이에 대한 최적의 역치를 정하는 것과 2) 통합의 수단으로 사용할 평균의 종류를 결정하는 두 가지의 최적화 과정이 진행되었다.

1.3.1 변이 수준 점수의 역치 최적화

유전자 내 유전적 다양성 (genetic variability)에 영향을 주는 중립 변이 (neutral variant)의 효과를 최소화 하기 위하여, 계산에 포함할 기능 변이에 대한 최적의 역치를 결정하는 변이 점수 최적화 과정이 수행되었다. 특히 약물 변이의 경우 전장 유전체 분석을 통하여 non-coding 영역에서 보고된 약물 – 변이 간 연관관계가 많이 보고 되었다. 합성 연관 (synthetic association)은 이러한 non-coding 영역의 변이들이 연관 관계 (association)에 있을 뿐 실제 약물 부작용의 원인 (causality)은 non-coding 변이와 연관 (linkage disequilibrium) 되어 있는 coding 영역의 여러 드문 변이들에 의해 설명된다는 이론으로, 해당 이론에 기반하여 PharmGKB 에 보고된 non-coding 약물 변이와 약물 간 연관 관계를 gold standard 로 두고 coding 변이로 계산한 GVB 값으로 gold standard 를 맞추는 데 최적의 성능을 갖는 임계값을 확인하였다 [17]. 결과적으로 SIFT 점수 기준의 10 개의 역치 값 중,

0.7 점 이하의 변이만을 포함하였을 때 인종에 관계 없이 가장 안정적인 성능을 보인다는 것을 확인하였다 (표 3). 인종 특이적 분석 결과는 PharmGKB 데이터베이스에 보고된 약물 – 변이 간 연관을 인종 별로 분류하여 구성한 gold standard 세트를 바탕으로 계산한 것으로, 인종 별 변이 분포를 고려하여 서로 다른 역치 값을 결정하는 데 도움이 될 수 있는 자료이나 사용한 gold standard 의 수가 매우 적어 불안정하다는 한계 때문에 참고자료로만 사용되었다. 최적화 된 역치 값의 기준을 만족하는 변이들만 포함했을 때 중립 변이에 의한 효과가 제거되는지 확인해 본 결과, 중립 변이의 경우 역치 값 적용 후 그 효과가 상당 부분 사라진다는 것을 확인할 수 있었다 (그림 1).

표 3. 최적의 역치를 결정하기 위한 변이 점수 최적화 분석 결과

	인종 특이적					인종 비특이적				
역치	Global	AFR	AMR	ASN	EUR	Global	AFR	AMR	ASN	EUR
0.1	0.703	0.804	0.675	0.678	0.676	0.659	0.646	0.647	0.684	0.654
0.2	0.704	0.806	0.679	0.670	0.679	0.655	0.641	0.643	0.679	0.653
0.3	0.699	0.803	0.675	0.662	0.675	0.648	0.635	0.636	0.671	0.645
0.4	0.703	0.799	0.681	0.668	0.683	0.657	0.643	0.649	0.676	0.657
0.5	0.706	0.797	0.690	0.660	0.692	0.663	0.653	0.654	0.678	0.662
0.6	0.710	0.799	0.698	0.662	0.700	0.666	0.653	0.659	0.68	0.666
0.7	**0.713**	**0.798**	**0.702**	**0.667**	**0.703**	**0.670**	**0.657**	**0.666**	**0.684**	**0.671**
0.8	**0.713**	0.797	0.701	0.667	0.702	**0.670**	0.656	0.665	0.684	0.670
0.9	0.712	0.797	0.701	0.666	0.702	**0.670**	0.656	0.666	0.683	0.671
1	0.712	0.797	0.701	0.666	0.702	**0.670**	0.656	0.666	0.683	0.670

각 cell 은 AUC 값을 의미함. AFR: African, AMR: American, ASN: Asian, EUR: European

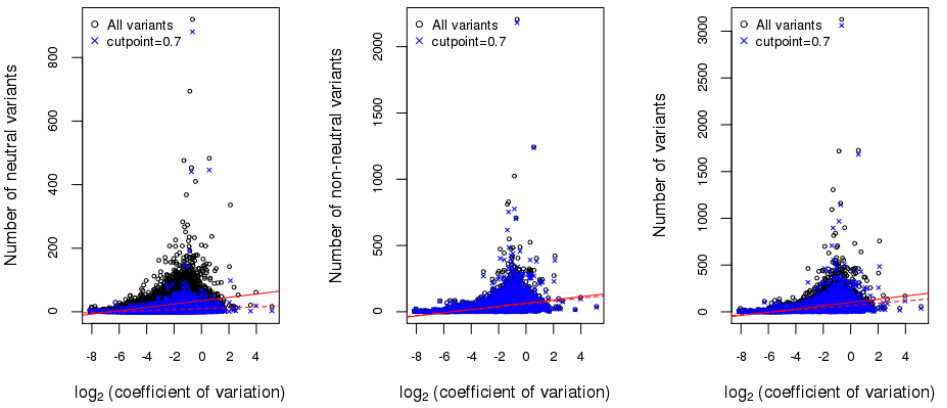


그림 1. 역치 적용 전 후의 중립 변이 제거 효과. Coefficient of variation 은 1000 genomes project 의 2504 명에 대한 유전자 별 사람 간 유전자 점수의 상대 표준 편차를 나타냄.

1.3.2 변이 수준 점수 통합 방법

유전자에서 발견된 변이의 영향을 효과적으로 통합하기 위하여 non-arithmetic Pythagorean means을 사용하였다. 산술 평균, 기하 평균, 조화 평균, 그리고 제곱의 방법을 사용하여 계산된 GVB 의 예측 정확도를 비교해보면 제곱 〉 조화 평균 〉 기하 평균 〉 산술 평균의 순서로 예측 정확도가 높은 것을 확인할 수 있다 (표 4) [17]. 그러나 제곱의 경우 계산적 이점은 존재할 수 있지만 실세계 모델링에서는 변이의 적응도 (fitness)를 최대화하는 방향으로 작용한다는 점에서 자연스럽지 않다고 판단하였다 [20]. 기하 평균은 n 개의 양수 값을 모두 곱한 것의 n 제곱근으로, 정규화된 결과를 평균화 할 때 결과 값이 기준 값에 대한

비율로 표시되는 특성을 가진다. 조화 평균은 주로 변화율을 확인하는 데 적합한 방법론으로, 필요에 따라 GVB 계산에 기하 평균 외 조화 평균을 사용할 수 있다. 적합한 평균의 종류를 결정하는 과정에 대한 평가는 추후 체계적으로 보완 할 필요성이 있다.

약물 유전체는 물론이고 희귀 질환, 복합 질환 등 다양한 유전적 배경에서 해당 방법론에 대한 유용성이 체계적으로 평가된 적이 없었다. 따라서 본 연구에서는 1) 소아 급성 림프모구성 백혈병 환자의 6-MP 약물 독성 예측에서 유전자 수준 변이 부담 점수의 임상성 유용성을 평가하고, 2) 약물 유전체 외 희귀 질환, 복합 질환에서의 유전자 수준 변이 부담 점수의 활용 가능성을 평가하며, 해당 질환들의 유전적 조성 차이를 반영하여 각 유전적 배경에 최적화된 유전자 수준 변이 부담 점수를 제안한다. 마지막으로 3) 탐색의 관점에서 유전자 수준 변이 부담 점수의 활용 가능성을 평가한다.

표 4. 평균 종류 별 유전자 수준 변이 부담 점수의 예측 정확도 비교.

인종\방법	산술평균	기하평균	조화평균	제곱
전체 (n=2504)	0.5935	0.6271	0.6363	0.6502
아프리카인(n=661)	0.5983	0.6261	0.6325	0.6502
미국인 (n=347)	0.5958	0.6314	0.6402	0.6491
동양인 (n=993)	0.5908	0.6266	0.6376	0.6520
유럽인 (n=503)	0.5912	0.6264	0.6360	0.6472

각 cell 은 AUC 값을 의미함.

2 소아급성 림프모구성 백혈병 환자의 6-MP 약물 독성 예측에서 유전자 수준 변이 부담 점수의 임상적 유용성 평가

2.1 연구배경

6-mercaptopurine (6-MP)은 소아급성 림프모구성 백혈병 (ALL, Acute lymphoblastic leukemia)의 유지요법에서 일반적으로 매일 사용되는 약물로, 해당 약물을 투여받은 일부 환자에서 중증 중성구 감소, 골수 기능 억제 및 간독성을 포함한 치명적인 약물 유발 부작용이 보고되어 각 환자의 상태에 따라 적절한 약물 용량을 결정하는 맞춤형 약물 투약 요법을 제공하는 것이 매우 중요하게 인식되어 왔다 [21].

초기 6-MP 복용량을 결정하는 가장 이상적인 방법은 6-MP 대사 산물의 농도를 모니터링하거나 체외 활성도 프로필 등을 직접 실험함으로써 약물 부작용을 가능성을 평가하는 것이지만 [22-26], 이러한 방법론은 시간이 오래 걸리고 동시에 높은 비용이 소모되는 비효율적인 구조를 가지고 있어서 임상 실무에 적용하는 것에는 한계가 있었다 [27].

그러나 최근 유전체 기술의 발달로 유전 변이와 6-MP 약물 독성 간의 관계가 입증됨에 따라, 개인의 유전적 특성에 따라 약물에 대한 반응성을 예측하는 방법론이 제시되었다. 치오퓨린 (Thiopurine) 계열

약물 독성에 가장 결정적인 역할을 한다고 알려진 유전자는 *TPMT* 로, 해당 유전자에 변이를 가지고 있는 사람의 경우 효소 활성이 감소하기 때문에 약물 용량을 조절하거나 혹은 약물 사용을 중지하는 등의 치료법 변경이 필요하였다 [28]. 그러나 *TPMT* 는 인종 간 유전적 차이가 매우 큰 유전자로 아시아인에서는 거의 변이가 발견되지 않기 때문에 한국인에서는 약물의 적정 용량 예측에 적용하기 어려웠다. 최근 치오퓨린계 약물 독성 관련 아시아인종에서 높은 빈도로 발생되는 새로운 예측 마커인 *NUDT15* 이 보고되면서 [29-31] CPIC (Clinical Pharmacogenetics Implementation Consortium)에서는 두 유전자에서 발견되는 변이와 그 유전형에 기반한 치오퓨린 약물 적정 용량 가이드라인을 제시하였다 [32]. 현재 CPIC 에서 약물의 적정 용량을 제시하는데 사용하는 방법론은 스타 대립 유전자 유전형에 기반한 것으로, haplotype block 내 위치하는 여러 변이의 조합에 따른 기능을 예측하게 된다 [33]. 그러나 해당 방법론은 임상에 적용하기에 여러 문제점이 발생할 수 있는데, 1) 명명법이 매우 복잡하다는 점, 2) 스타 대립 유전자가 부여되었으나 기능은 밝혀지지 않은 경우가 많아 실제 약물 독성 예측에 활용되는 allele 은 극히 일부라는 점, 3) 아직 스타 대립 유전자가 부여되지 않은 매우 드물게 발견되거나 혹은 개인이 특이적으로 가지는 (private) 변이의 효과가 반영되지 못한다는 점, 4)

따라서 인종간 혹은 개인간 유전적 조성의 차이가 큰 약물 유전체적 특성에 따라 기존에 연구된 적 있는 인구 집단 외에서는 그 활용이 제한된다는 점이 바로 그것이다.

차세대 염기 서열 분석법의 시대에 유전자 내 완전한 변이 다양성을 예측하는 것이 가능해짐에 따라, 본 연구에서는 흔한 변이 뿐 아니라 드물거나 혹은 개인에게서만 발견되는 변이의 효과까지 모두 통합할 수 있는 유전자 수준의 변이 부담 (GVB) 점수를 제안한다. GVB 점수가 변이에 따른 효소 활성도 예측에 어느정도 기여할 수 있는지 그 유용성을 평가하기 위하여, 244 명의 ALL 환자에서의 예측대비 실제 투약 용량 (DIP)을 임상적 최종 평가 변수로 두고 스타 대립 유전자 기반의 약물 적정 용량 예측 결과와 정확도를 비교한다. 최종적으로는 스타 대립 유전자 기반 약물 독성 예측 방법론에 대한 대안적 방법론으로서의 가치를 평가한다.

2.2 재료 및 방법론

2.2.1 환자군 설정과 임상 데이터 수집

유지요법 중 6-MP 약물을 투약 받은 한국인 소아급성 림프모구성 백혈병 환자 298 명이 2 개 병원 [아산 (AMC, Asan Medical Center), 서울대 (SNUH, Seoul National University Hostital) 병원] 으로부터 모집되었다. 이 중 재발, 줄기 세포 이식, 버킷 임파종, 혼합된 표현형의 급성 백혈병, 유아기 ALL, 혹은 VHR (very high risk) 등의 제외 사유를 가지는 환자를 제외한 총 244 명의 환자를 최종 연구 대상으로 설정하였다. 모든 환자는 동의서를 작성하였으며, 본 연구는 양 병원의 연구 윤리 위원회로부터 승인받았다. 12 주 약물 사이클 동안 6-MP 의 용량 (per meter body surfate)이 기록되었으며, 마지막 사이클에서의 약물 용량이 독성 반응률을 초과하지 않는 최대 허용 용량이라고 판단하였다. 두 병원의 환자들은 최소 500 에서 최대 1500/ μL 의 타겟 절대 호중구수를 유지하기 위하여 같은 치료 요법과 약물 조절 가이드라인을 사용하여 치료 받았다. 유전자형에 기반한 약물 용량 조절은 수행되지 않았다 (표 5). 최종 244 명의 환자에 대해 Ion AmpliSeq^{TM} Exome panel 을 이용하여 엑솜 시퀀싱을 수행하였다.

표 5. 환자의 임상적 특성.

특성	연구 코호트		
	AMC	**SNUH**	전체
환자 수	95	149	244
진단시 나이 (년), mean±sd[†]	5.23 ± 1.8	8.57 ± 4.6	7.26 ± 4.1
성별			
남	52	93	145
여	43	56	99
마지막 사이클 6-MP 약물 용량 **(mg/m2/day),** mean ± sd (N)			
6-MP < 12.5	8.14 ± 1.7 (2)	6.25 ± 2.9 (4)	6.88 ± 2.6 (6)
12.5 ≤ 6-MP < 25	17.39 ± 3.4 (4)	19.40 ± 3.6 (9)	18.78 ± 3.7 (13)
25 ≤ 6-MP < 37.5	32.19 ± 3.4 (10)	30.72 ± 4.0 (16)	31.28 ± 3.8 (26)
37.5 ≤ 6-MP < 50	44.52 ± 3.7 (13)	45.80 ± 3.5 (14)	45.18 ± 3.6 (27)
6-MP ≥ 50	79.15 ± 18.1 (66)	78.84 ± 23.1 (106)	78.96 ± 21.3 (172)
Total	65.37 ± 26.6 (95)	65.03 ± 30.0 (95)	65.16 ± 28.7 (244)

[†]진단시 나이는 한 명에 대한 데이터가 존재하지 않음. 6-MP: 6-Mercaptopurine, AMC: Asan Medical Center, SNUH: Seoul National University Hospital

2.2.2 유전자 단위의 변이 부담 점수 계산

NUDT15 과 *TPMT* 에 대한 유전자 단위의 변이 부담 점수를 계산하였다. SIFT 〈 0.7 의 점수를 가지는 변이를 기능적 위해도를 보일 가능성이 있을 것으로 가정하였으며, SIFT 점수가 부여되지 않은 삽입 혹은 결실 변이는 단일 변이보다 위해도가 클 것으로 판단하고 $1e^{-08}$ 의 점수를 부여하였다.

$$\boldsymbol{G_i = \{v \mid v \text{ with a SIFT score less than } 0.7\}}$$

유전자량 효과를 반영하여 양쪽 allele 모두에 변이를 가지는 homozygous 형태보다 한쪽 allele 에만 변이를 가지는 heterozygous 형태의 변이의 심각도가 낮을 수 있도록 가중치를 반영하였다.

$$_{adj}\boldsymbol{v_j} = \begin{cases} (\boldsymbol{SIFT\ score})^{\boldsymbol{0.5}}, & \boldsymbol{if\ v_j \in G_i\ and\ heterozygote} \\ \boldsymbol{SIFT\ score}\ , & \boldsymbol{if\ v_j \in G_i\ and\ homozygote} \end{cases}$$

n 개의 유해한 변이를 가지는 유전자 G_i 각각에 대하여, n 개 변이에 대한 SIFT 점수의 기하 평균을 계산하여 한 유전자 내에 존재하는 모든 변이의 누적 효과를 통합하는 유전자 단위의 변이 부담 점수를 산출하였다.

$$GVB(G_i) = \begin{cases} 1 & , if\ n(G_i) = 0 \\ \left(\prod_{j=1}^{n} adj\, v_j \right)^{\frac{1}{n}} & , if\ n(G_i) > 0 \end{cases}$$

유전자 내 한개의 유해한 변이도 발견되지 않는 경우에는 GVB 로 1 이 부여되었다. *NUDT15* 과 *TPMT* 각각에 대한 유전자 점수의 기하 평균을 통해 6-MP 독성 민감도를 예측하는 약물 점수가 계산되었다. 유전자 단위의 변이 부담 점수는 0 에서 1 사이의 범위를 가지며, 0 에 가까울 수록 그 위해도가 높은 것으로 판단한다.

2.2.3 스타 대립 유전자 추론 및 분자 표현형 변환

244 명의 ALL 환자를 약물 독성 위험도에 따라 세 군의 분자 표현형 군으로 분류하기 위한 목적으로 PHASE 2.1.1 도구를 사용하여 일배체형을 추정하였다 [34, 35]. 추정된 일배체형 정보에 기반하여 PharmGKB 의 스타 대립 유전자 정의 테이블로부터 일치하는 스타 대립 유전자를 추출하였다 [36]. 각 일배체형으로 부터 추출된 두 스타 대립 유전자의 조합을 통해 개인 별 스타 대립 유전자 유전형을 산출하였으며, Moriyama *et al.* (*NUDT15*)과 PharmGKB 에서 제공하는 유전형-표현형 변환 테이블을 통해 약물에 대한 독성군을 예측하였다

[37].

2.2.4 진단적 정확도 예측

예측 정확도를 평가하기 위하여 마지막 사이클에서의 예측대비 실제 투약 용량을 나타내는 DIP 를 6-MP 약물 독성에 대한 지표로 두고 GVB 와 스타 대립 유전자 기반 방법론 간 예측 결과를 비교하였다 [38]. 이분법적 분류 모델 하에서 약물에 대한 고위험 군을 크게 9 가지 cut-off 레벨 (*i.e.*, 5%, 10%, 15%, 25%, 35%, 45%, 60%, 80%, and 100%)로 나누고, 각 역치에서의 예측 성능을 pROC 패키지를 이용한 ROC 분석을 통해 확인하였다 [39]. 더불어, 임상에서 고위험군 예측의 지표로 사용되는 DIP 〈 25% 기준을 만족하는 환자를 고위험군으로 두었을 때, 각 방법론의 민감도, 특이도, 양성 예측도, 음성 예측도, 그리고 정확도를 비교 평가하여 그 임상적 유용성을 확인하였다. 모든 통계적 분석은 R 3.5.1 버전을 사용하여 수행되었다.

2.3 결과

2.3.1 유전자 단위의 변이 부담 점수와 스타 대립 유전자 기반 방법론 간 연관성

엑솜 염기 서열 시퀀싱 (whole exome sequencing, WXS) 방법을 통해 얻어진 유전체 정보로 부터 각 개체의 *NUDT15* 과 *TPMT* 에 대한 haplotype block 을 유추하였으며, 이와 매치되는 스타 대립 유전자 유전형을 추출하였다. 각 스타 대립 유전자 유전형은 최종적으로 3 가지 종류 (Poor, Intermediate, 그리고 Normal metabolizer; 각각 PM, IM, 그리고 NM)의 분자 표현형으로 변환되었다. 244 명의 ALL 환자의 유전체 정보로부터는 두 유전자에서 총 10 가지 종류의 스타 대립 유전자가 발견되었다 (표 6).

244 명의 ALL 환자에서의 세 종류의 분자 표현형 분포를 확인해보면 (표 7), *NUDT15* 의 경우 49 명 (20.1%) 에서 약물 독성의 가능성이 있는 환자들이 발견 된 반면 *TPMT* 의 경우 7 명 (2.9%)에서만 약물 독성의 가능성이 있는 환자들이 발견되었다. 이는 두 유전자가 인종 특이적인 유전 변이 분포를 보임을 시사한다. *NUDT15* 에 변이를 가지지 않는 NM 군의 경우 해당 유전자에 변이를 가지는 PM 또는 IM 군에 비해 평균적으로 유의하게 높은 DIP 를 가지고 있어 (NM= 67.608 ±

28.2, n = 195; IM= 56.452 ± 28.2, n=48; PM=5.172, *n*=1), 스타 대립 유전자 기반의 방법론이 약물 적정 용량 그룹을 성공적으로 분류하고 있음을 확인할 수 있다 (그림 2). *TPMT* 의 경우 NM 군이 IM 군에 비해 평균적으로 높은 DIP 를 가지는 분포는 일치하였으나 (NM= 65.702 ± 28.4, n = 237; IM= 46.805 ± 35.7, n=7), 한국인에서 발견되는 변이 빈도가 매우 낮기 때문에 군 간 유의한 차이를 보이지는 않았다 (Mann-Whitney U test p=0.101).

표 6. 기능이 알려진 allele 로 정의한 244 명의 소아 백혈병 환자에 대한 매치된 스타 대립 유전자.

유전자	발견된 allele의 갯수	244명에서 발견된 allele	빈도 (%)
NUDT15	6	*1	438 (89.75)
		*2	6 (1.23)
		*3	35 (7.17)
		*4	4 (0.82)
		*5	4 (0.82)
		*6	1 (0.20)
TPMT	4	*1	127 (26.02)
		*1S	354 (72.54)
		*3C	6 (1.23)
		*6	1 (0.20)

일배체형은 PHASE2 를 통해 유추함. 스타 대립 유전자는 PharmGKB 일배체형 변환 표를 사용하여 지정함.

표 7. 예측된 효소 대사 표현형의 분포.

분자 표현형	기능	*NUDT15*	*TPMT*
Poor (%)	No function| No function	1 (0.41)	NA
Intermediate (%)	Normal | No function	48(19.67)	6 (2.46)
	Normal | Decreased	NA	1 (0.41)
Normal (%)	Normal | Normal	195 (79.92)	237 (97.13)
전체 (%)		244 (100)	244 (100)

분자 표현형은 PharmGKB 유전자형-표현형 변환 표와 Moriyama *et al.* (*NUDT15*), the CPIC guideline (*TPMT*)을 사용하여 매치함.

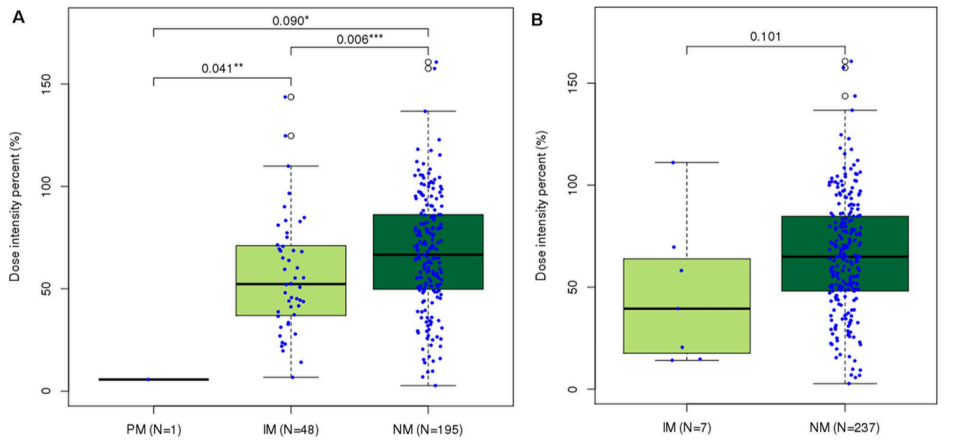


그림 2. 소아 백혈병 환자에서 스타 대립 유전자 기반 분자 표현형 그룹에 따른 6-MP 의 마지막주기 용량 강도 백분율의 분포. 분자 표현형 그룹 별 (A) *NUDT15* and (B) *TPMT* 에서의 용량 강도 백분율 분포. *NUDT15* 의 Normal metabolizers 는 intermediate (p=0.006) 와 poor (p=0.090) 그룹 보다 유의하게 높은 용량 강도를 보임. Mann-Whitney U test, * $p < 0.1$, ** $p<0.05$, and *** $p<0.01$.

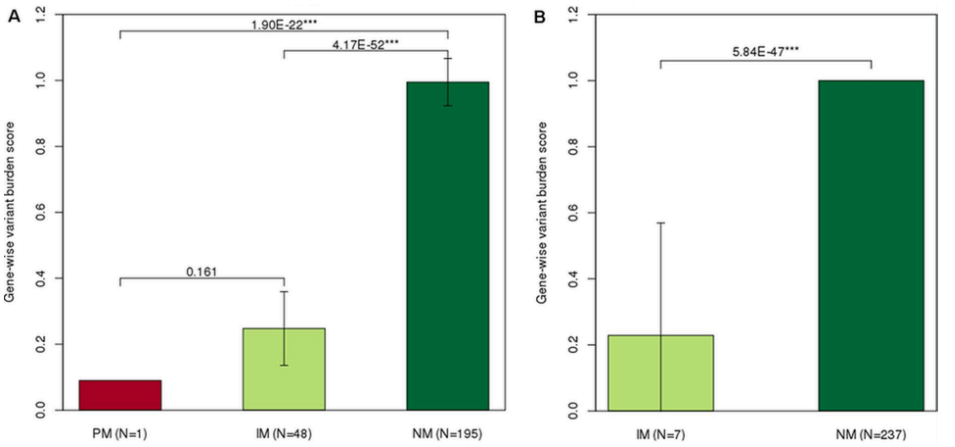


그림 3. 스타 대립 유전자 기반 분자 표현형 그룹에 따른 유전자 수준 변이 부담 점수의 분포. (A) *NUDT15* 과 (B) *TPMT* 분자 표현형 그룹의 유전자 수준 변이 부담 점수. Normal metabolizers 는 intermediate (*NUDT15* p=4.17E-52; *TPMT* p=5.84E-47) 와 poor (*NUDT15* p=1.9E-22) 그룹 보다 유의하게 높은 약물 용량 분포를 보임. Mann-Whitney U test, * $p < 0.1$, ** $p<0.05$, and *** $p<0.01$.

스타 대립 유전자 기반으로 분류된 세 분자 표현형 그룹 (PM, IM, 그리고 NM) 별 유전자 단위의 변이 부담 점수의 분포를 확인해보면, 약물 독성에 대한 심각도가 높을 것으로 예측되는 PM 군으로 갈 수록 낮은 GVB 점수를 보이고, 반대로 약물 독성에 대한 심각도가 높지 않을 것으로 예측되는 NM 군으로 갈 수록 높은 GVB 점수를 보이는 트렌드가 유지되어 (*NUDT15*: PM = 0.09, IM = 0.248 ± 0.1, 그리고 NM = 0.995 ± 0.1; *TPMT*: IM=0.229±0.3,NM=1± 0.0), 두 방법론 간 양의 상관 관계를 가지고 있음을 확인하였다 (그림 3).

마지막으로 GVB 점수에 따라 DIP 의 분포가 어떻게 달라지는 지 그 관계를 확인해보면, *NUDT15* 의 경우 244 명 중 GVB 점수가 높은 사람이 평균적으로 높은 DIP 를 보이는 유의한 양의 상관관계를 가졌다 (그림 4A, Kruskal–Wallis test p = 0.016, Spearman's rank correlation p = 0.001 (ρ = 0.21), Kendall's rank correlation p = 0.001 (τ = 0.17)). 이 중 공변량 효과를 제거하기 위하여 *TPMT* 에 변이를 가지는 2 명의 환자를 제외하고 군 간 연관성을 재 확인해보면 유의성이 조금 더 증가하는 것을 확인할 수 있었다. *TPMT* 의 경우 낮은 점수의 GVB 를 가지는 환자에서 낮은 DIP 가 요구되는 양의 방향의 트렌드는 여전히 유지되었으나, 변이 빈도가 매우 낮아 거의 대부분의 환자가 WT 군

(GVB=1)으로 분류되었기 때문에 유의한 연관성을 보이지는 않았다 (그림 4B , t-test p = 0.408, Spearman's rank correlation p = 0.272 (ρ = 0.07), Kendall's rank correlation p = 0.271 (τ = 0.06)). 해당 분석은 트렌드를 확인하기 위하여 코호트에서 발견된 모든 유전자 점수에 대한 DIP 분포를 그래프로 나타내었으나, 각 bin 에 포함되는 샘플 수가 너무 적어 분포를 확인하는 데 한계가 있을 수 있다.

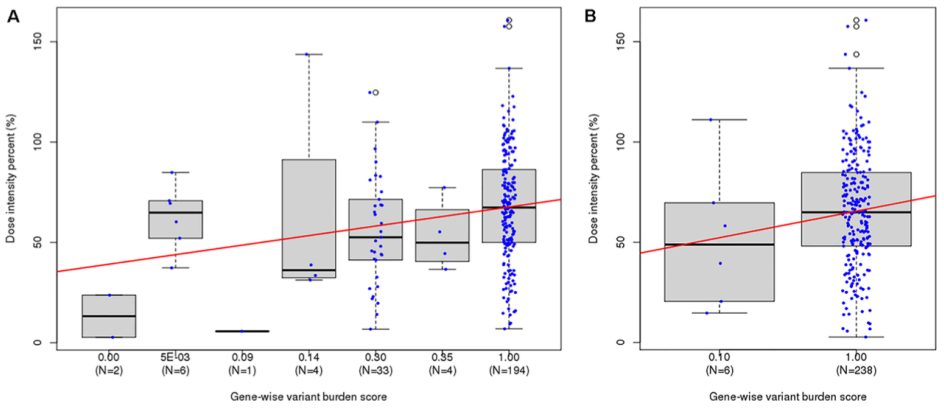


그림 4. 유전자 수준 변이 부담 점수에 따른 6-MP 의 마지막 주기 용량 강도 백분율의 분포. (A) GVB^{NUDT15} (Kruskal-Wallis p-value = 0.016, Spearman's rank correlation p-value = 0.001 (ρ=0.21), Kendall's rank correlation p-value=0.001 (τ=0.17)) (B) GVB^{TPMT} (Kruskal-Wallis p-value = 0.271, Spearman's rank correlation p-value = 0.272 (ρ=0.07), Kendall's rank correlation p-value=0.271 (τ=0.06)).

2.3.2 유전자 단위의 변이 부담 점수와 스타 대립 유전자 기반 방법론 간 약물 독성 군 예측 성능의 비교

예측 성능을 평가하기 위하여 9 개의 서로 다른 cut-off 레벨 (*i.e.*, DIP < 5%, 10%, 15%, 25%, 35%, 45%, 60%, 80%, and 100%)에서 6-MP 고위험군을 정의하고, GVB 가 각 레벨의 위험군을 얼마나 잘 예측하는지 ROC 분석을 통해 확인하였다. *NUDT15* 의 경우 약물 독성의 심각도가 높을 것으로 예상되는 군에서 상대적으로 더 높은 예측 성능을 보였으며 (그림 5A, AUC=0.998 (DIP < 5%), 0.676 (DIP < 10%), 0.639 (DIP < 15%), 그리고 0.627 (DIP < 25%)), 일반적으로 PM 군을 분류하는 DIP < 25%의 기준에서 기존의 스타 대립 유전자 기반의 예측법 (AUC = 0.618) 보다 미세하게 높은 예측 성능을 보였다. 이 중 *TPMT* 에 변이를 가지는 2 명의 환자를 제외하면 더욱 향상된 예측 성능을 보였다 [그림 5B, AUC=0.998 (DIP < 5%), 0.676 (DIP < 10%), 0.669 (DIP < 15%), 그리고 0.653 (DIP < 25%)]. *TPMT* 의 경우 낮은 변이 빈도로 GVB 와 스타 대립 유전자 기반 방법론 모두 DIP 예측 성능이 매우 낮았다 (그림 5C-D). 그러나 DeLong's 테스트 결과 모두 군 간 통계적으로 유의한 차이를 보이지는 않았다.

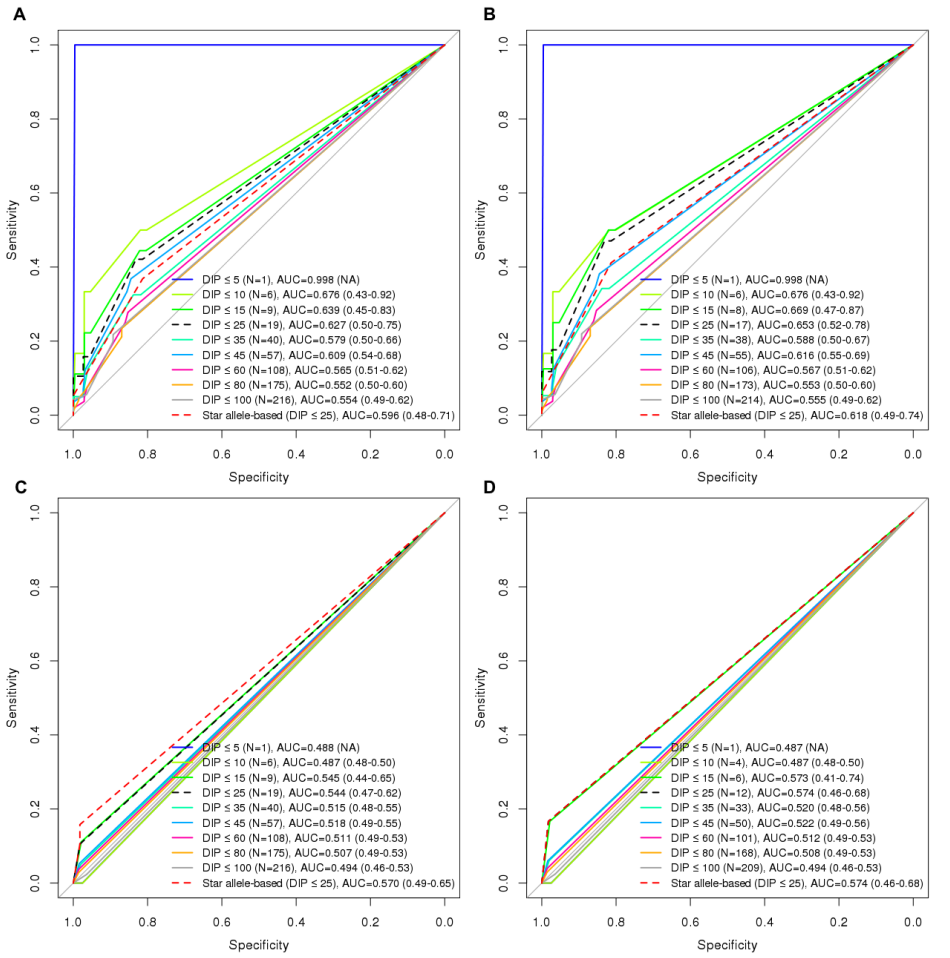


그림 5. 소아 희귀암 6-MP 에 대한 대립 유전자 기반 분자 표현형과 유전자 수준 변이 부담 점수 사이의 진단 정확도 비교. AUROC 분석을 통해 계산된 진단 정확도. (A) *TPMT* 변이를 가지고 있는 두 명을 제외한 GVB^{NUDT15} (DeLong's p-value=0.163), (B) GVB^{NUDT15} (DeLong's p-value=0.163), (C) *NUDT15* 변이를 가지고 있는 7 명을 제외한 GVB^{TPMT} (DeLong's p-value=0.5), 그리고 (D) GVB^{TPMT} (DeLong's p-value=0.841). AUC 의 괄호 안에는 95% 신뢰 구간을 표시함. DIP: Dose Intensity Percent, AUC: Area Under the Receiver Operating Curve.

더 중요하게는, GVB 가 정량적 특성을 갖는 점수 체계라는 점에 입각하여 두 유전자의 효과를 하나의 점수로 합친 $GVB^{NUDT15,TPMT}$ 의 예측 성능이 각각의 유전자 점수보다 뛰어났다는 것이다 (그림 6). 특히 임상적으로 약물 독성의 심각도가 높을 것으로 분류하는 기준인 DIP 〈 25%에서 0.677로 각 유전자의 GVB나 스타 대립 유전자 기반 방법론과 비교하였을 때 가장 좋은 예측 성능을 보였다는 점에서 기존의 방법론과 견줄만한 단순하면서도 직관적인 방법론임을 입증하였다. 다만, 해당 분석에서도 방법론 간 통계적으로 유의한 차이를 보이지는 않았다.

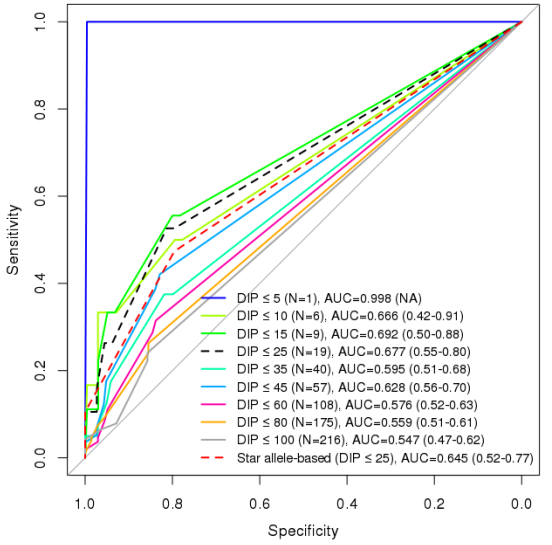


그림 6. *NUDT15* 과 *TPMT* 의 효과를 혼합했을 때 스타 대립 유전자 기반 분자 표현형과 유전자 수준 변이 부담 점수의 6-MP 불내성 예측 진단 정확도의 비교. AUC 분석을 통해 계산한 $GVB^{NUDT15,TPMT}$ 의 진단 예측도(DeLong's p-value=0.175). AUC 의 괄호 안에는 95% 신뢰 구간을 표시함. DIP: Dose Intensity Percent, AUC: Area Under the Receiver Operating Curve.

2.3.3 유전자 단위의 변이 부담 점수와 스타 대립 유전자 기반 방법론 간 약물 독성 군 예측 정확도의 비교

6-MP 적정 용량 제시에 있어서 GVB 의 임상적 효용성을 평가하기 위하여, GV 에 의해 분류된 위험군에 대한 진단적 정확도를 계산하고 그 결과를 기존의 스타 대립 유전자 기반 분류 결과와 비교하였다 (표 8). GVB 방법론은 기존의 스타 대립 유전자 기반 방법론에 비해 1 명의 고위험 군 (DIP 〈 25)과 4 명의 저위험 군 (DIP 〉 25)을 더 잘 예측하였으며, 결과적으로 민감도 (47.36% 에서 52.63%), 특이도 (79.56% 에서 81.33%), 양성 예측도 (16.36% 에서 19.23%), 음성 예측도 (94.70% 에서 95.31%), 그리고 정확도 (77.05% 에서 79.10%)의 모든 카테고리에서 향상된 예측 정확도를 보였다. 본 연구에서 제시한 결과들을 종합했을 때, 소아 ALL 환자에서 6-MP 약물 독성군을 예측하는 데 있어서 *in silico* 예측 방법을 활용하여 계산된 GVB 점수는 기존의 경험적으로 획득한 스타 대립 유전자 기반 방법론보다 향상되거나 적어도 비교할만한 예측력을 가진다는 것을 확인할 수 있다.

표 8. 소아 희귀암 환자에서 스타 대립 유전자 기반 분자 표현형과 유전자 수준 변이 부담 점수간 6-MP 독성 예측 결과의 비교. (A) 스타 대립 유전자 기반 분자 표현형과 (B) 유전자 수준 변이 부담 점수에 대한 진단 정확도 결과표.

(A)

	DIP			
NUDT15 와 *TPMT* 분자 표현형	≤ 25	〉25	전체	
PM+IM	9	46	55	양성예측도 16.36% (9/55)
NM	10	179	189	음성예측도 94.70% (179/189)
전체	19	225	244	
	민감도 47.36% (9/19)	특이도 79.56% (179/225)		정확도 77.05% (188/244)

(B)

	DIP			
유전자 수준 변이 부담 점수	≤ 25	〉25	전체	
GVB ^{NUDT15,TPMT} ≤ 0.3	10	42	52	양성예측도 19.23% (10/52)
GVB ^{NUDT15,TPMT} 〉0.3	9	183	192	음성예측도 95.31% (183/192)
전체	19	225	244	
	민감도 52.63% (10/19)	특이도 81.33% (183/225)		정확도 79.10% (193/244)

2.5 고찰

최근 6-MP 약물 독성과 연관된 다양한 종류의 새로운 유전 변이들이 잇달아 보고됨에 따라, 환자 개개인의 상황을 고려하여 적정 용량의 약물을 예측하는 것이 핵심적인 문제로 자리잡았다 [40]. 여러 유전 변이의 상호 작용을 고려하였을 때 변이 하나로 개인의 약물 독성을 예측하는 것은 매우 믿기 어렵기 때문에, 여러 변이의 효과를 통합할 수 있는 새로운 방법론의 개발이 요구되어왔다. 본 연구에서는 소아 ALL 환자에서 6-MP 약물 독성을 예측하는 데 있어서 GVB 점수의 임상적 유용성을 현재 임상에서 가장 많이 사용되는 스타 대립 유전자 기반 방법론과 비교 평가하였다. 최근 업데이트 된 CPIC 가이드라인을 확인해보면 새롭게 발견되는 변이들에 대해 지속적으로 새로운 스타 대립 유전자를 부여하고 있으나, 보고된 변이들의 인종 특이성 때문에 1) 실제 연구에 사용된 인종을 제외하고는 해당 변이가 전혀 발견되지 않거나 (반대로 연구되지 않은 인종에서 나타나는 변이들의 효과는 전혀 반영할 수 없거나), 2) 혹은 해당 변이가 발견된다고 하더라도 그 기능이 입증되지 않은 경우가 많고, 3) 한 allele 의 기능을 입증하기 까지 매우 오랜 시간이 소요된다는 점에서 그 유용성에 한계가 있다. 이러한 측면에서 GVB 점수는 기존의 스타 대립 유전자 기반 방법론과

비교하였을 때 몇 가지 이점이 존재하는데, 1) 기존 방법론이 범주형 결과를 제시하는 반면에 GVB 는 정량적 수치로 제공한다는 점; 2) 각 유전자의 개인 간 유전적 조성 차이를 측정할 수 있다는 점; 3) 흔한 변이 뿐 아니라 드물거나 개인에게서 새롭게 발견되는 변이의 효과까지 통합하여 하나의 점수로 제공한다는 점; 4) 기존에 연구된 인종 혹은 집단에 의존하지 않는 독립적인 방법론이라는 점; 5) 다양한 유전자 간 상호 작용이나 아직 발견되지 않은 새로운 유전자를 탐색하기 위한 체계적인 분석 방법론으로서 활용될 수 있다는 점; 6) 이론상 모든 코딩 변이에 대한 위해도를 제공하는 *in silico* 예측 방법의 특성상 기존에 보고되지 않은 변이의 효과까지도 체계적으로 적용할 수 있다는 점이 바로 그것이다.

특히 GVB 점수 방법론에서는 고위험군으로 분류하였으나 스타 대립 유전자 기반 방법론에서는 저위험군으로 분류하였던 1 명의 환자는 23.7%의 낮은 DIP 를 보여 실제 약물 독성의 위험이 있었는데, 해당 환자가 보유하고 있었던 새로운 결실 변이는 CPIC 가이드 라인에 불명확한 기능을 가지는 *NUDT15**9 으로 등록되었고, 2019 년 업데이트된 CPIC 가이드라인에서는 기능 상실 allele 로 변경되었다 [41]. 즉 GVB 는 아직 보고되지 않은 새로운 변이 효과를 체계적으로 통합 적용함으로써 이보다 앞서 해당 결과를 예측하였음을 시사한다. 더불어

NUDT15 과 *TPMT* 두 유전자에 동시에 변이를 가지고 있었던 1 명의 환자는 스타 대립 유전자 기반 분류법에서는 각각 중간 정도의 약물 독성 (IM)을 보이는 것으로 판정하였지만 실제로는 다른 환자에 비해 매우 심각한 약물 독성을 보였는데 (DIP=14%), GVB 점수의 경우 두 유전자 효과를 모두 고려하여 통합한 하나의 정량적 수치를 제공할 수 있었다는 점에서 고위험군 예측에 이점을 보였다.

그러나 본 연구에는 몇 가지 제한점이 존재한다. GVB 점수 계산에 사용되는 핵심 자원인 다양한 *in silico* 기반 예측 점수의 대부분은 단일 변이에 대한 위해도만을 제공하기 때문에 삽입 혹은 결실 변이는 무조건 기능 상실이 있을 것으로 판단하고 그 효과를 임의로 반영하였다. 그러나 짧은 삽입 혹은 결실 변이에 대한 효과까지 제공하는 CADD 점수에서 삽입 혹은 결실 변이의 예측 효과를 확인해보면 단일 변이 대비 위해도가 반드시 높은 것은 아니기 때문에, 임상적 활용을 위해서는 약물 민감도에 대한 해당 변이들의 영향에 대한 실험적 검증이 반드시 필요하다 [42]. 더불어 희귀 질환인 소아 ALL 환자군의 특성상 제한된 샘플 내에서 평가가 이루어졌지만, GVB 점수 방법론의 타당성을 평가하기 위해서는 다양한 인종을 대상으로 독립 연구 군에서 좀 더 확장된 약물-유전자에 대해 평가해보는 과정이 반드시 필요하다.

3 유전자 수준의 변이 부담 점수: 약물, 복합질환, 그리고 희귀질환 연관 유전자에 대한 유전적 특성화

3.1 연구배경

차세대 시퀀싱 기술이 발전하면서, 임상적으로 의미있는 다양한 종류의 유전 변이가 기여하는 바를 예측하고 수집하는 것이 가능해졌다. 단일 변이를 기준으로 기능적 영향을 추정하는 가장 강력한 접근법 중 하나로 '점수 기반 시스템'이 도입되었다. 양성 (benign) 변이로 부터 유해한 (pathogenic) 변이를 구분하기 위해 백개 이상의 다양한 *in silico* 예측 도구가 제안되었다 [43]. 그러나, 단일 변이를 기반으로 하는 해석은 생물학적 의미를 얻기에는 한계가 있으며 [44], 서로 다른 점수 체계를 사용하는 경우에 같은 변이에 대해서 서로 상충하는 결과를 내는 경우도 있기 때문에 혼란이 있었다 [45–47].

이러한 변이 별 접근 방식의 한계를 완화하고자 최근 유전자 혹은 영역 기반의 접근법이 잠재적 위해성을 예측하는 방법으로 대두되었다. 특히 다양한 유전자 단위 점수를 바탕으로 하는 연구들이 활발히 진행되고 있는데, 예를 들어, 유전자 단위의 점수 중 pLI 및 RVIS 는 유전 변이에 대한 감내성 (genic intolerance)의 개념을 바탕으로 특정 인종에서 예측 대비 실제 발견된 기능 변이의 갯수를 유전자

단위로 측정하는 방법론 [9, 13], GDI 점수는 직관적으로 유전자 내 발견된 모든 변이의 부하 정도 (loads)를 계산하는 방법론으로 [10], 다양한 연구에서 질병과 연관될 것으로 예측되는 후보 유전자를 추출하는 데 이러한 유전자 단위 점수를 사용해왔다. 현재의 유전자 단위 점수들은 주로 희귀 질환 연구에 초점을 두고 사용되어 왔는데, 약물 반응성 관련 혹은 복합 질환과 관련된 유전자는 희귀 질환 유전자와는 그 특성이 매우 다를 것이기 때문에 각 질환과 관련된 유전자의 특성에 대해 체계적으로 분석하고 그 특성에 맞는 새로운 개념의 유전자 점수를 개발하는 것에 대한 필요성이 대두되었다.

희귀 질환은 주로 이른 나이에 생명을 위협할 만큼 심각도가 높은 형태의 질병이 발생하는 반면, 복합 질환은 늦은 나이에 환경적, 다양한 외부적 요인에 의하여 만성적으로 발생한다는 특성이 있다 [48]. 약물 부작용은 정의상 특정 약물에 노출 되기 전까지는 어떠한 증상도 보이지 않는다는 점에서 일반적인 질환과는 그 발생 과정에 차이가 있다. 특히 유전적인 관점에서, 희귀 질환은 주로 유전자 내 존재하는 매우 소수의, 드물지만 영향력이 큰, 그리고 강력한 진화적 압력 하에 있는 변이에 의해 발생하지만, 복합 질환은 상대적으로 흔하고, 다양한 유전자에 흩어져있는, 약한 음성 선택 (negative selection) 하에 존재하는 변이와 연관되어있다고 알려져 왔다 [49, 50]. 약물 변이는 질병

변이와는 다른 맥락에서 발생하는데, 주로 인종 간 빈도 차이가 큰 한 개 혹은 그 이상의 흔한 변이와 드문 변이로 이루어진 일배체형 (haplotype) 단위에 기인한다는 데 차이가 있다 [51, 52]. 심지어는, 유전 변이를 설명하는 용어 선택에 있어서도 구분하려는 노력이 이루어져 왔다. ClinVar 에서는 [53] 희귀 질환과 연관된 유전 변이를 지칭하는 용어로 흔히 "병원성의" (pathogenic), "병원성의 가능성이 있는" (likely pathogenic), 그리고 "양성의" (benign) 등을 사용해왔다. 그러나 약물 변이를 지칭하는 용어로는 "대사: 빠른, 중간의, 느린" (metabolism: rapid, intermediate, poor), "효능: 저항성, 반응성, 민감성" (efficacy: resistant, responsive, and sensitive), 또는 "위험한" (risk) 등으로 구분하여 사용하는 것을 권장하고 있다. 이렇게 상이한 특성에 따라 질병 유전자와 약물 유전자를 구분하여 접근하는 것이 반드시 필요함에도 불구하고, 현재의 유전자 단위 접근법에서는 서로 다른 유전적 카테고리를 구분하여 접근하려는 시도가 이루어지지 않고 있다.

무엇보다 pLI, RVIS 및 GDI는 특정 인종 내에서 발견되는 유전 변이의 빈도를 바탕으로 집단 단위의 개념으로 계산되기 때문에, 계산된 점수는 특정 인종의 유전적 조성에 매우 의존하게 된다는 특성이 있다. 그러나 이러한 집단 단위의 점수는 약물 유전학에서 특정 약물에 대한 개인의 반응성 차이를 설명하는 가장 강력한 예측 요소인 개인간,

그리고 인종간 차이를 반영하지 못한다 [14]. 따라서 특정 인종의 유전적 조성에 치우치지 않는, 독립적인 유전자 단위 점수를 개발하는 것에 대한 필요성이 대두되었다. 게다가, 기존 점수의 특성과 한계를 정확히 이해하고 후보 유전자를 추출했다고 하더라도, 최종적으로 해당 유전자의 기능이나 관련 질병에 대한 정보를 얻는 과정에서 사용하는 생물학적 경로 (biological pathway) 혹은 유전자 온톨로지 (gene ontology) 에서 지식 중심의 편견 (knowledge-driven bias)이 발생한다는 것은 후보 유전자 추출 과정에서 매우 잘 알려진 문제이다 [54]. 이러한 문제를 완화하는 한 가지 방법은 질병 군 간 대표되는 유전자들이 가지는 차별화된 유전적 특성을 수집하고, 여기서 얻은 패턴을 추후 분석 과정에서 아직 많이 연구되지 않은, 혹은 기능이 잘 알려지지 않은 유전자를 포함한 모든 유전자에 편견 없이 체계적으로 적용하는 것이다.

유전자 수준의 변이 부담 (GVB) 점수는 개인 별로 한 유전자 내 발생한 모든 위해 변이의 누적 효과를 통합한 유전자 단위의 점수로, 이미 다양한 약물 유전체 연구들에서 그 쓰임이 입증되었다. 본 연구에서는 약물 유전자 뿐 아니라 희귀, 복합 질환에서 유전자 점수의 유용성을 평가한다. 더불어, 일곱 가지 분자 유전적 특성 (number of paralogs, number of singletons, per-person mutability, protein-protein

interaction (PPI) degree, coding sequence (CDS) length, McDonald–Kreitman neutrality index (NI), and protein complexity)을 바탕으로 약물 및 희귀, 복합질환 연관 유전자를 구분할 수 있는 특징적 패턴을 확인한다. 마지막으로, 각 카테고리에서 얻은 특성을 적용한 보정된 GVB (adjusted GVB)의 예측 성능을 평가함으로써 카테고리 별로 최적화된 알고리듬을 확인한다.

3.2 재료 및 방법론

3.2.1 GVB 계산

1000 Genomes Project 에서 제공하는 Phase 3 데이터를 사용하여 발견된 변이에 대해 주석 달기를 진행하였다 [55]. 단백질 코딩 유전자 영역은 ANNOVAR 를 사용하여 결정되었다 [56]. 변이의 위해도는 일곱가지 *in silico* 예측 방법을 사용하여 예측되었다: SIFT (http://sift.jcvi.org/) [57–59], CADD [42], PolyPhen2 HIVD, PolyPhen2 HVAR [60], PhyloP [61], MutationTaster [62], 그리고 GERP++ [63]. 2504 명의 개인 유전체 데이터 각각에 대하여 17,502 개의 유전자에 대한 GVB 점수를 계산하였다 (그림 7).

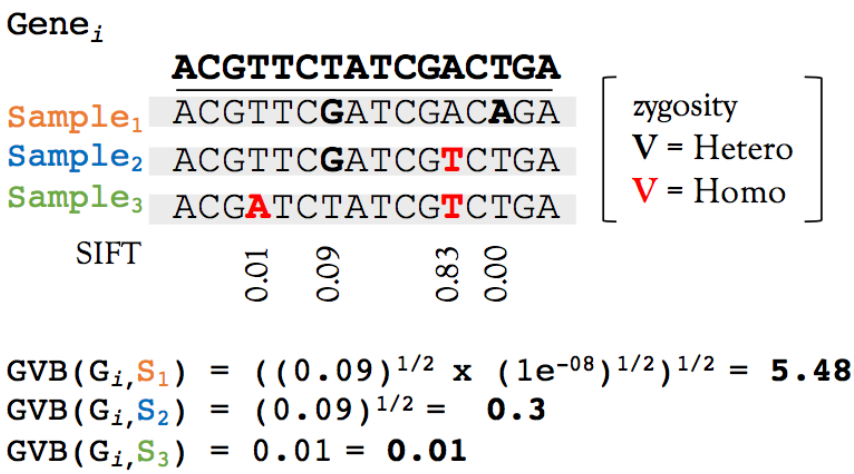


그림 7. GVB 점수 계산 흐름의 요약. *i* 번째 유전자 (Gene$_i$)의 GVB 는 세 명의 사람 (S1, S2, and S3) 각각에 대해 계산된다. 각 변이는 유전형에 따라 서로 다른 가중치를 부여하여 동형접합 변이를 가지는 경우 더 위해한 효과를 가지도록 한다. 0.7 점 이하의 SIFT 점수를 가지는 경우에 아미노산 변경에 영향을 주는 기능 변이라고 판단하고 GVB 점수 계산에 포함한다. GVB 계산 결과에 따르면, Gene$_i$는 S2 보다 S1 에 더 위해한 영향을 가질 수 있다.

3.2.2 포괄적인 유전 카테고리에 대한 유전자 목록 수집

유전자 수준의 점수 방법론이 다양한 유전적 배경에서 얼마나 관련 유전자를 잘 예측하는 지 평가하기 위하여, 세 가지 유전자 카테고리에 속하는 알려진 유전자 목록을 준비하였다: 약물 연관 표현형, 복합 질환, 그리고 희귀 질환. 약물 카테고리의 경우, 다섯 가지 약물 유전자 세트를 Absorption, Distribution, Metabolism, and Excretion (ADME) (http://www.pharmaadme.org) 와 PharmGKB (4 월 22, 2020) [64]; ADME core, ADME extended, ADME all, PharmGKB VIP, and

PharmGKB 로부터 추출하였다. 충분히 포괄적인 평가를 하기 위해서, 다음의 약학적 특성에 따라 유전자를 분류하였다; PharmGKB 로부터 약물학적 효과 (*i.e.*, Toxicity, Metabolism/PK, Dosage, and Efficacy), DrugBank (http://www.drugbank.ca/; v5.15) [65] 로부터 약동학적 (PD, pharmacodynamic) 그리고 약력학적 (PK, pharmacokinetic) 표현형 (*i.e.*, Target, Enzyme, Transporter, and Carrier), 그리고 효소 family (*i.e.*, Cytochromes P450 (CYP) 그리고 UDP-glucuronosyltransferase (UGT)) (http://www.genenames.org/cgi-bin/genefamilies).

복합 질환의 경우, GAD (7 월 23, 2011) [66] 로부터 다음의 총 16 개의 질환군 목록을 추출하였다; Aging, Cancer, Cardiovascular, Chemdependency, Developmental, Hematological, Immune, Infection, Metabolic, Neurological, Normalvariation, Pharmacogenomic, Psych, Renal, Reproduction, 그리고 Vision. 유전자와 표현형 간 유의한 상관관계를 나타내는 "Y" 주석이 달린 항목만을 사용하였다. 전체 19 개의 질환 목록 중, 불분명한 표현형 (unknown 과 others 카테고리)을 나타내거나 항목에 너무 적은 수의 유전자를 포함하고 있는 경우 (mitochondrial 카테고리)를 제외하고 사용하였다.

희귀 질환 카테고리의 경우, OMIM [67] 으로부터 추출하여 Petrovski *et al.*의 논문에 사용한 총 6 개 목록을 (*i.e.*, Recessive,

Haploinsufficiency, Dominant-negative, De novo, OMIM all, 그리고 Non-disease) 그대로 사용하였다. 다섯개의 생존 불가능한 (nonviable) 유전자 목록은 Bartha *et al.* (*i.e.*, *In vivo* essential, *In vitro* essential, Mice essential) 그리고 Petrovski *et al.* (*i.e.*, Mouse Genome Informatics (MGI) lethality 그리고 MGI seizure)의 작업으로 부터 추출하였다 [9, 68]. 유전자 점수의 성능은 R 소프트웨어의 pROC 패키지를 사용하여 평가하였다 [39]. 예측 성능은 모든 가능한 점수를 임계값으로 하여, 알려진 약물 또는 질병 유전자 세트 중 임계값보다 작은 점수를 가진 유전자의 비율을 계산하는 방법으로 측정되었다.

3.2.3 유전자 특이적인 분자 유전적 특성

유전자에 일곱개의 생물학적으로 의미있는 분자 유전 특성들 (molecular genetic features)의 주석을 달았다. paralog 의 갯수와 CDS 길이, 선택적 압력의 척도인 NI 와 단백질 복잡도의 척도인 D 값은 Itan *et al.*의 작업으로부터 추출하였다 [10]. Itan *et al.*의 작업에 의하면, CDS 길이와 paralog 의 갯수는 Ensembl Biomart (version 75)로부터 추출하였다. 인간에서의 단백질-단백질 상호작용 정도 (PPI degree)는 IntAct 데이터베이스 (4 월 22, 2020)로부터 12,128 유전자에 대하여 추출하였다 [69]. Singleton 의 갯수와 per-person mutability 는 1000 Genomes Project

데이터로부터 계산하였다. 해당 데이터에서 한번만 발견된 코딩 변이가 singleton 으로 정의되었다. 각 유전자에 대한 per-person mutability 는 $M_i = (\sum_{j=1}^{n} V_{ij})/n$ 로 정의되었는데, 이 때 V_{ij} 는 j번째 사람의 i 유전자 내 SIFT 주석이 달린 모든 변이를 의미한다.

3.2.4 분자 유전 특성을 사용한 GVB 점수 보정

일곱개의 파라미터를 가중치로 사용하여 GVB 점수를 보정하였다. 일곱 가지 생물학적으로 의미있는 분자 유전 특징을 나누거나 곱한 네 가지의 변형된 수식을 통해 보정된 GVB 점수를 생성하였으며 (표 9), 각각의 보정된 GVB 의 예측 성능을 다양한 유전자 카테고리를 바탕으로 체계적으로 평가하였다. 파라미터 값 중 0 이 있는 경우, 값이 0 으로 수렴하는 현상을 막기 위하여 10^{-8} 의 값을 부여하였다. GVB 계산에 사용된 일곱 개의 서로 다른 *in silico* 점수 중 SIFT 가 가장 좋은 예측 성능을 보였기 때문에, 이후의 분석은 SIFT 기반의 GVB 점수를 사용한 결과를 기준으로 작성되었다.

표 9. GVB* 보정 수식.

1. $G_{1/f}(G_i) = \frac{\left(\prod_{j=1}^{n} v_j\right)^{\frac{1}{n}}}{f_i}$	2. $G_{1/f2}(G_i) = \frac{\left(\prod_{j=1}^{n} v_j\right)^{\frac{1}{n}}}{{f_i}^2}$
3. $G_f(G_i) = \left(\prod_{j=1}^{n} v_j\right)^{\frac{1}{n}} \times f_i$	4. $G_{f2}(G_i) = \left(\prod_{j=1}^{n} v_j\right)^{\frac{1}{n}} \times {f_i}^2$

3.3 결과

3.3.1 GVB, RVIS, pLI, 그리고 GDI 점수의 특징 비교

pLI, RVIS 및 GDI 방법론은 주어진 인구집단 내 변이의 상대 빈도(다양성)와 변이 위해도 점수를 사용한다 (표 10). 해당 점수들은 인구집단 단위로 유전자 별 점수를 제공하는 반면에, GVB 는 인구집단에 의존하지 않고 개인 단위로 유전자 별 점수를 제공한다. 유전자 별 인구집단 단위의 GVB 점수는 해당 인구집단 내 속하는 개인 유전자 점수를 통합함으로써 쉽게 생성할 수 있다. 즉, GVB 는 개인은 물론이고 인구집단 특이적인 유전자 단위 점수를 동시에 제공한다. pLI, RVIS 및 GDI 는 서로 다른 인구 집단의 유전적 조성에 민감하지만, GVB 는 이와 독립적인 특성을 갖는다. 더불어, GVB 는 목적에 따라 다양한 *in silico* 예측 점수를 활용하여 계산할 수 있으며, 유전적으로 중요한 의미를 갖는 분자적 요소들을 유연하게 고려할 수 있다. 이렇게 분자적 유전 요소를 고려한 GVB 점수를 GVB*라고 명명하였다.

표 10. 유전자 수준 우선순위 방법론의 특성 비교.

특성	GVB	RVIS	pLI	GDI
점수 할당 기준	사람별	인구집단별	인구집단별	인구집단별
인구집단 의존성	독립적	ESP6500	ExAC	The 1000 Genomes Project, Phase 3
주요 변인	SIFT, CADD, PolyPhen HIVD, PolyPhen HVAR, PhyloP, MutationTaster, 그리고 GERP 등의 *in silico* 예측 점수	변이 빈도	Loss of function 변이 빈도	변이 빈도와 CADD 점수
점수 범위	[0, 1]	[-8.29, 29.75]	[0,1]	[0, 42.91]
표준화	Y	N	N	N
점수 유형	절대치	상대치	상대치	상대치

3.3.2 다양한 유전적 카테고리에서 GVB 의 예측 성능 평가

그림 8 은 다양한 약물, 복합 질환, 희귀 질환의 주요/하위 카테고리에서 GVB, GDI, RVIS, 그리고 pLI 의 연관 유전자 판별력을 나타내는 AUC (areas under the receiver operating characteristic curves) 값을 나타낸 것이다. 전반적으로, GVB 는 약물 유전자에서 pLI, RVIS, GDI 대비 높은 판별력을 보이는 것을 확인할 수 있다 (상위 패널). 복합 질환에서도 거의 대부분의 하위 카테고리에서 GVB 는 다른 점수 체계 대비 좋은 판별력을 보였다 (가운데 패널). Paralog 의 갯수와 CDS 길이를 적용하여 보정한 GVB*는 거의 대부분의 카테고리에서 GVB 대비 향상된 예측 성능을 보였다.

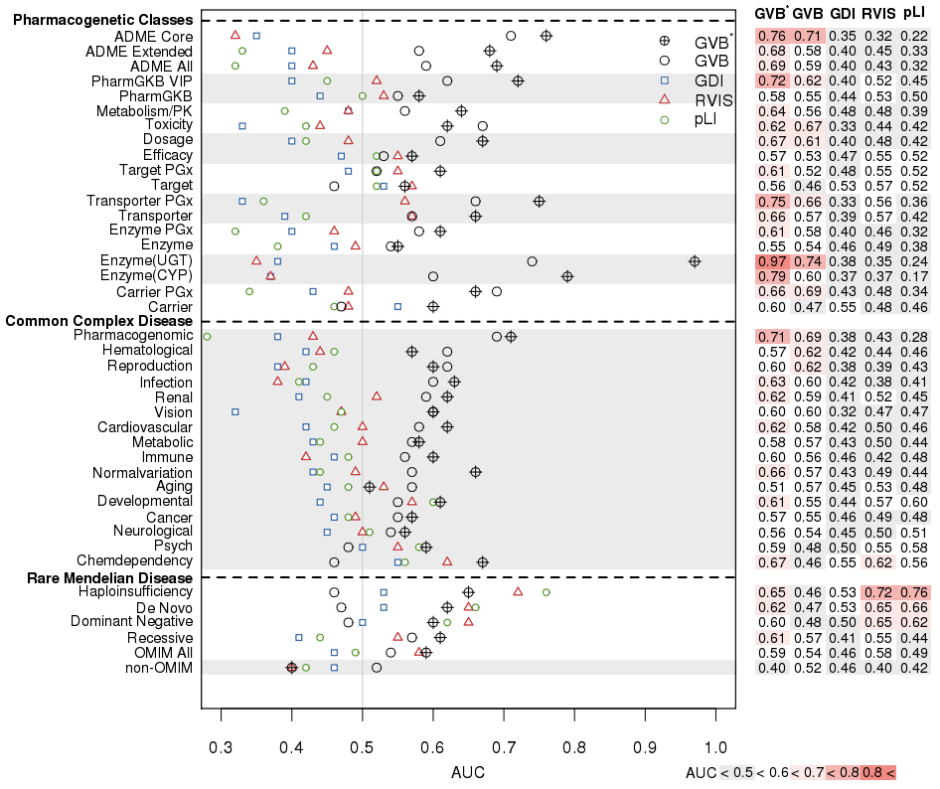


그림 8. GVB, RVIS, pLI, 그리고 GDI 의 약물, 복합질환, 희귀질환 유전자 예측에 대한 성능 비교. 각 하위 범주의 AUC 값은 강도 스케일에 따라 회색 (낮은 AUC)에서 빨간색 (높은 AUC) 까지의 색상으로 표시됨. GVB* 는 paralogs 의 수와 CDS 길이로 보정됨. GAD, Genetic Association Database; ADME, Absorption, Distribution, Metabolism, and Excretion; PK, pharmacokinetic; PGx, pharmacogenetic; OMIM, Online Mendelian Inheritance in Man.

GVB 와 GVB*는 모든 주요 약물 하위 카테고리에서 높은 AUC 를 보였다 (그림 9A). 약물 카테고리 내에서도, 주요 (core) 약물 카테고리에서의 예측 성능이 주변 (peripheral) 카테고리에 비해 높았는데, ADME core 〉 ADME extended 의 순서 (AUC_{GVB}: 0.71 〉 0.58; AUC_{GVB*}: 0.76 〉 0.68), 그리고 PharmGKB VIP 〉PharmGKB 의 순서 (AUC_{GVB}: 0.62 〉 0.55; AUC_{GVB*}: 0.72 〉 0.58)로 높은 AUC 값을 보이는 것을 확인할 수 있었다. 특히 두 점수 모두 약력학 (PK, Pharmacokinetic)적 요소인 효소와 transporter, carrier 카테고리에서 약동학 (PD, Pharmacodynamic)적 요소인 target 카테고리 보다 더 잘 동작하였다.

복합 질환 카테고리의 경우, 각 방법론이 높은 예측력을 보이는 카테고리가 따로 존재하였다 (그림 9A). GVB 는 앞선 결과와 동일하게 약물유전체 카테고리에서 높은 예측력 (AUC_{GVB*}: 0.71, AUC_{GVB}: 0.69)을 보인 반면, GDI 는 약물 의존성 (AUC_{GDI}: 0.55)과 정신 질환 (AUC_{GDI}: 0.50)에서, pLI 와 RVIS 는 약물 의존성 (AUC_{pLI} ,AUC_{RVIS}: 0.56, 0.62)과 발달 질환 (AUC_{pLI} ,AUC_{RVIS}: 0.60, 0.57)에서 각각 가장 높은 예측력을 보였다. 그러나 대부분의 복합 질환 하위 카테고리에서 GVB 가 다른 방법론에 비해 높은 예측 성능을 보였다.

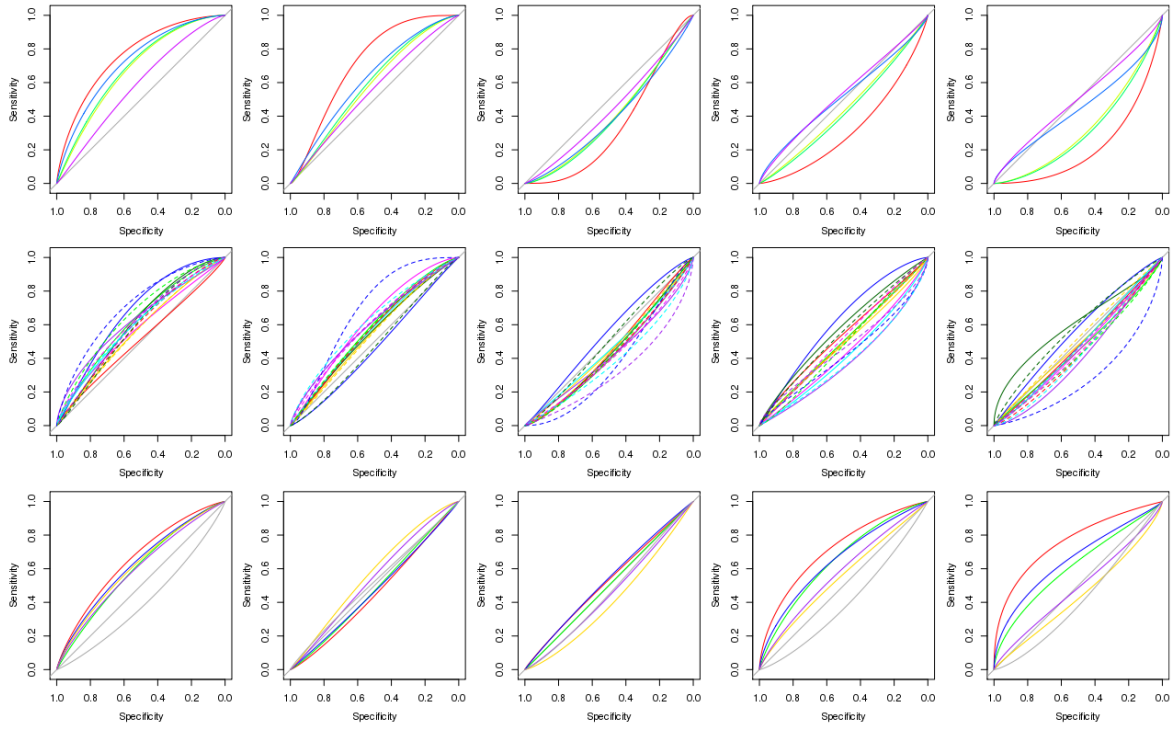
희귀 질환 카테고리에서는 pLI, RVIS 와 GDI 가 Haploinsufficiency, De novo, 그리고 Dominant negative 의 카테고리에서 GVB 보다 뛰어난 성능을 보였다. 특히 pLI 가 희귀 질환 카테고리에서 Haploinsufficiency 〉 De novo 〉 Dominant negative 의 순서로 가장 높은 성능을 자랑했다 (0.76 〉 0.66 〉 0.62). GVB 는 희귀 질환 하위 카테고리 중 Recessive 에서 가장 높은 예측 성능을 보였는데 (AUC_{GVB}: 0.57), 이는 GVB 가 다른 세 점수와는 상호 보완적인 역할을 할 수 있다는 것을 시사한다. CDS 길이에 의해 보완된 GVB*는 질병 심각도에 따라 Haploinsufficiency 〉 De novo 〉 Dominant negative 〉 Recessive 로 높은 예측 성능을 보였다.

GVB 와 다른 세 방법론은 모든 카테고리에서 서로 역방향의 예측 순위를 보였는데, 이는 인구 집단 내 발견된 변이들에 대한 감내성 (intolerance)에 초점을 두는 pLI, RVIS, 및 GDI 의 방법론이 위해도와 유전적 다양성에 초점을 두는 GVB 와 다른 관점을 가지기 때문으로 해석된다. 다양한 *in silico* 예측 점수를 바탕으로 계산한 GVB 점수에서도 이러한 경향성은 그대로 유지되었는데, 이는 매우 안정적인 (robust) 결과임을 의미한다.

추가로 하위 약물 카테고리에서의 예측 성능을 평가해보면 (그림 9B), GVB 는 약물 독성 관련 카테고리 (AUC_{GVB}: 0.66), 그리고 PK 유전자에서 가장 뛰어난 예측 성능 (Transporter PGx 〉 Enzyme PGx,

0.66 〉 0.58)을 보이며, paralog 의 갯수로 보정한 GVB*에서는 상대적으로 paralog 의 갯수가 많은 Metabolism/PK (AUC_{GVB*}: 0.64)와 Dosage (AUC_{GVB*}: 0.67), 그리고 Transporter PGx (AUC_{GVB*}: 0.75) 에 대한 예측 성능이 증가한 것을 확인할 수 있다. 특히 효소 관련 family 유전자의 경우 UGT family 에서 가장 좋은 성능을 (AUC_{GVB*}: 0.97, AUC_{GVB}: 0.74), CYP family (AUC_{GVB*}: 0.79, AUC_{GVB}: 0.60) 에서도 여전히 높은 성능을 보였다. 약물 카테고리에서도 여전히 GVB 와 다른 세 방법론 간 역방향의 예측 순위가 유지되었다.

(A)



	N	GVB*	GVB	GDI	RVIS	pLI
ADME Core	32	0.76	0.71	0.35	0.32	0.22
ADME Extended	267	0.68	0.58	0.40	0.45	0.33
ADME All	299	0.69	0.59	0.40	0.43	0.32
PharmGKB VIP	66	0.72	0.62	0.40	0.52	0.45
PharmGKB	1665	0.58	0.55	0.44	0.53	0.50

	N	GVB*	GVB	GDI	RVIS	pLI
Pharmacogenomic	19	0.71	0.69	0.38	0.43	0.28
Hematological	40	0.57	0.62	0.42	0.44	0.46
Reproduction	79	0.60	0.62	0.38	0.39	0.43
Infection	99	0.63	0.60	0.42	0.38	0.41
Renal	72	0.62	0.59	0.41	0.52	0.45
Vision	34	0.60	0.60	0.32	0.47	0.47
Cardiovascular	286	0.62	0.58	0.42	0.50	0.46
Metabolic	370	0.58	0.57	0.43	0.50	0.44
Immune	313	0.60	0.56	0.46	0.42	0.48
Normalvariation	22	0.66	0.57	0.43	0.49	0.44
Aging	31	0.51	0.57	0.45	0.53	0.48
Developmental	54	0.61	0.55	0.44	0.57	0.60
Cancer	275	0.57	0.55	0.46	0.49	0.48
Neurological	167	0.56	0.54	0.45	0.50	0.51
Psych	184	0.59	0.48	0.50	0.55	0.58
Chemdependency	54	0.67	0.46	0.55	0.62	0.56

	N	GVB*	GVB	GDI	RVIS	pLI
Haploinsufficiency	175	0.65	0.46	0.53	0.72	0.76
De Novo	467	0.62	0.47	0.53	0.65	0.66
Dominant Negative	364	0.60	0.48	0.50	0.65	0.62
Recessive	817	0.61	0.57	0.41	0.55	0.44
OMIM All	2131	0.59	0.54	0.46	0.58	0.49
Non-OMIM	14241	0.40	0.52	0.46	0.40	0.42

(B)

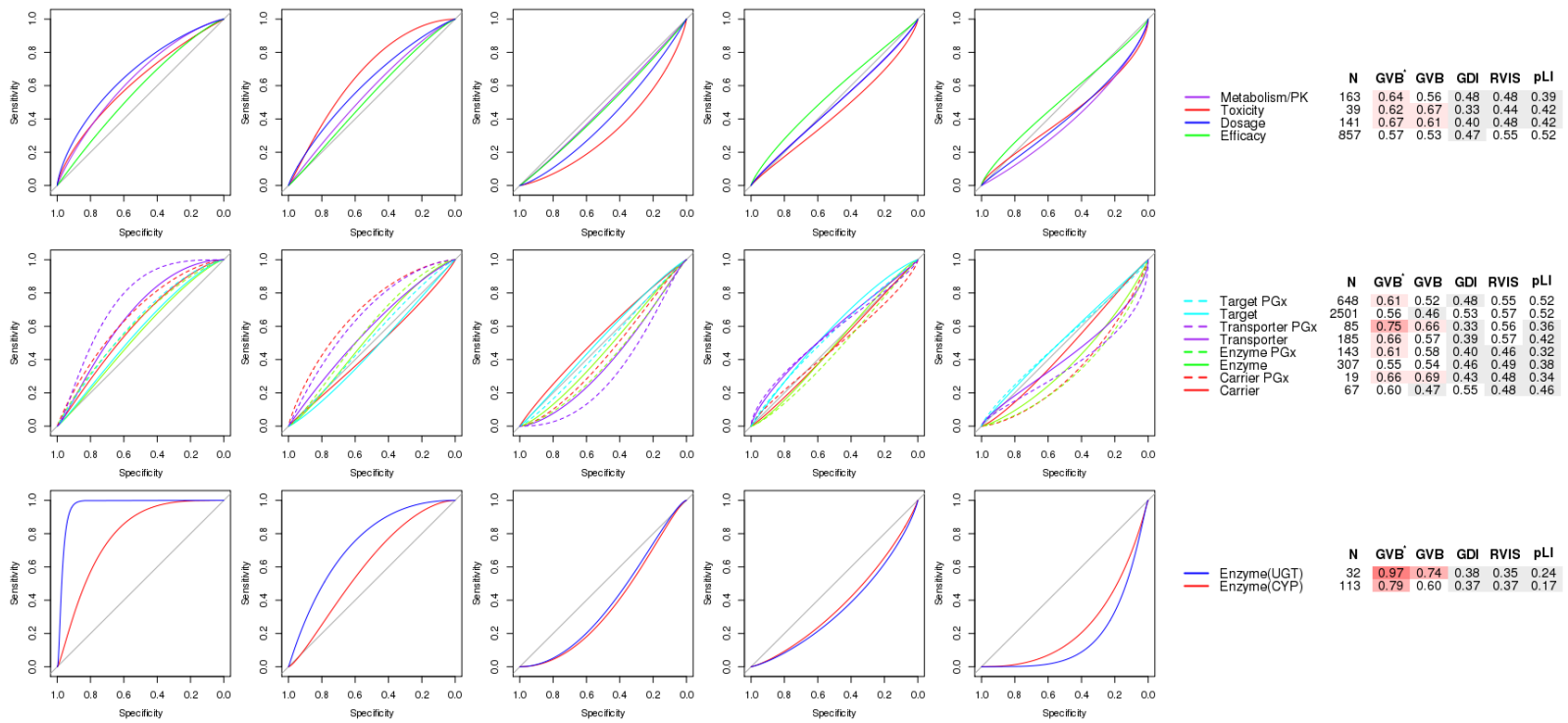


그림 9. 약물, 복합- 및 희귀질환 유전자 범주 및 하위 범주를 결정하기 위한 GVB, RVIS, pLI 및 GDI 의 성능 비교. (*A*) 다섯개의 약물, 16 개의 복합질환, 6 개의 희귀질환 (B) 네 개의 약물 효과, 네 개의 약리학 및 약동학적 표현형, 두개의 효소 집단에 대한 유전자 카테고리에서의 ROC curves. GVB*는 paralogs 의 갯수와 CDS length 로 보정됨.

3.3.3 약물, 복합질환, 희귀질환 유전자에 대한 유전적 특성화

한 유전자의 paralog 갯수는 계통 발생학적 측면에서 필수성 (essentiality)에 대한 지표를 나타낸다 [70]. Paralog 의 갯수가 상대적으로 많은 유전자는 기능적 파괴의 가능성에 대비하기 위하여 중복되는 기능을 가졌음을 의미하고, 반대로 paralog 의 갯수가 상대적으로 적은 유전자는 대체불가능한 필수적 역할을 하는 것으로 판단된다 [71, 72]. 희귀 질환 유전자와 비생존 (nonviable) 유전자는 강한 선택적 압력하에서 복합유전자 대비 적은 수의 paralog 갯수를 가진다 (그림 10). 반면, 약물 유전자는 질병 유전자 대비 많은 수의 paralog 를 보유한다. 비질병 (non-disease) 유전자의 경우 가작 적은 수의 paralog 를 보유하였다.

Singleton 이란 인구집단 내 단 한번 발견된 매우 드문 변이를 의미한다. 초기 연구에서 singleton 은 기술적 오류로 판단되기도 하였으나, 최근 singleton 변이의 기능적 중요도가 높아짐에 따라 해당 변이에 초점을 둔 연구들이 많아지고 있다. 본 연구에서는 한 유전자 내 발견된 singleton 의 갯수를 유전적 다양성 (genetic diversity)에 대한 지표라고 정의하였다. 희귀 질환 유전자는 다른 카테고리 대비 singleton 의 갯수가 유의하게 많았는데, 이는 진화적 측면에서 비교적

최근에 발생한 드문 변이들이 많이 존재한다는 것을 의미한다. 반대로 복합 질환 유전자에서는 가장 적은 singleton 갯수가 발견되었고, 이는 진화적으로 충분히 오랜 시간 전에 생겨나 인구집단 간 공유될 수 있었던 흔한 변이의 비중이 더 높게 나타난다는 것을 의미한다. 약물 유전자에서는 복합 질환과 희귀질환의 중간 정도 되는 singleton 갯수가 발견되었는데, 약물 유전 변이의 국소적 적응성 선택 (local adaptive selection)이 가능한 설명일 수 있다. 즉, 양성으로 선택된 변이가 유전적 히치하이킹 (genetic hitchhiking)에 의해 연결된 유해 변이의 빈도를 증가시키는 역할을 하게되고, 해당 인구 집단에서의 부작용을 회피하기 위한 이러한 과정이 결론적으로는 인구 집단 간 높은 유전적 차이를 보이게 하는 것이다.

한 유전자의 “개인 별 변이성” (per-person mutability)이란 한 개인에서 유전자 내 발생하는 단백질 코딩 변이를 견딜 수 있는 정도를 의미한다. 개인 별 변이성은 복합 질환에서 가장 높고 희귀 질환에서 가장 낮았는데, 이는 복합 질환 유전자에서는 유전자 내 다양한 변이가 발생하는 것을 용인할 수 있지만 희귀질환 유전자의 경우 변이 발생에 상대적으로 민감하게 반응한다는 것으로 해석할 수 있다.

단백질 상호 작용 정도 (protein-protein interaction degree)는 유전자의 기능적 중요도를 나타내는 가장 대표적인 척도 중 하나이다.

핵심적인 (core) 역할을 하는 유전자일 수록 단백질 상호 작용 정도가 높은데, 희귀 질환 유전자에서 복합 질환 유전자 대비 높은 단백질 상호 작용 정도를 보였다 [73-75]. 약물 유전자는 질환 유전자 대비 단백질 상호 작용의 정도가 낮았는데, 이는 약물 같은 외부 작용 없이는 증상을 일으키지 않는 약물 유전자의 특성을 고려했을 때 충분히 가능한 결과임을 예측할 수 있다.

Coding sequence (CDS)의 길이는 중요한 진화적 증거 중 하나로, 길이가 긴 유전자 일수록 복제 과정에서 많은 노력이 들어 느리게 진화하기 때문에 과발현시 해로울 수 있다는 보고가 있다. 희귀 질환 유전자의 CDS 길이는 다른 카테고리의 유전자 보다 유의하게 길었는데, 이는 기존의 연구에서도 반복되어 보고된 현상으로 희귀 질환 유전자를 구분하는 중요한 특성 중 하나로 분류될 수 있다.

중립 지수 (neutrality index)는 유전자에 대한 선택적 압력 (selective pressure)을 추정하기 위한 것으로, 1 보다 큰 경우 양성 선택 (positive selection)의 신호를 (an excess of nonsynonymous polymorphism), 1 보다 작은 경우 음성 선택 (purifying selection)의 신호를 (an excess of nonsynonymous divergence) 나타낸다 [76, 77]. 복합 질환의 평균 중립 지수는 희귀 질환의 평균 중립 지수 보다 유의하게 높았는데, 이는 복합 질환 유전자의 경우 희귀 질환 유전자

대비 짧은 CDS 길이에도 불구하고 아미노산 변화를 일으키는 많은 변이들이 누적되어 있음을 시사한다. 약물 유전자에서도 희귀 질환 유전자 대비 국소적 적응 선택되었던 많은 유해 변이들이 발견되는 것을 확인할 수 있다.

마지막으로, 단백질 복잡성 (protein complexity)은 Clark's distance 를 사용하여 추정한 아미노산 구성의 무질서도를 평가하는 척도로, 무질서한 영역에는 흔한 변이일 수록 축적되고 유해한 변이일 수록 고갈된다는 보고들이 있었다 [78, 79]. 우리 연구에서는 복합 질환 유전자에서 희귀 질환 대비 높은 D 값을 가졌는데, 이는 앞서 보고된 여러 연구들과 일치하는 방향성을 갖는다 [80].

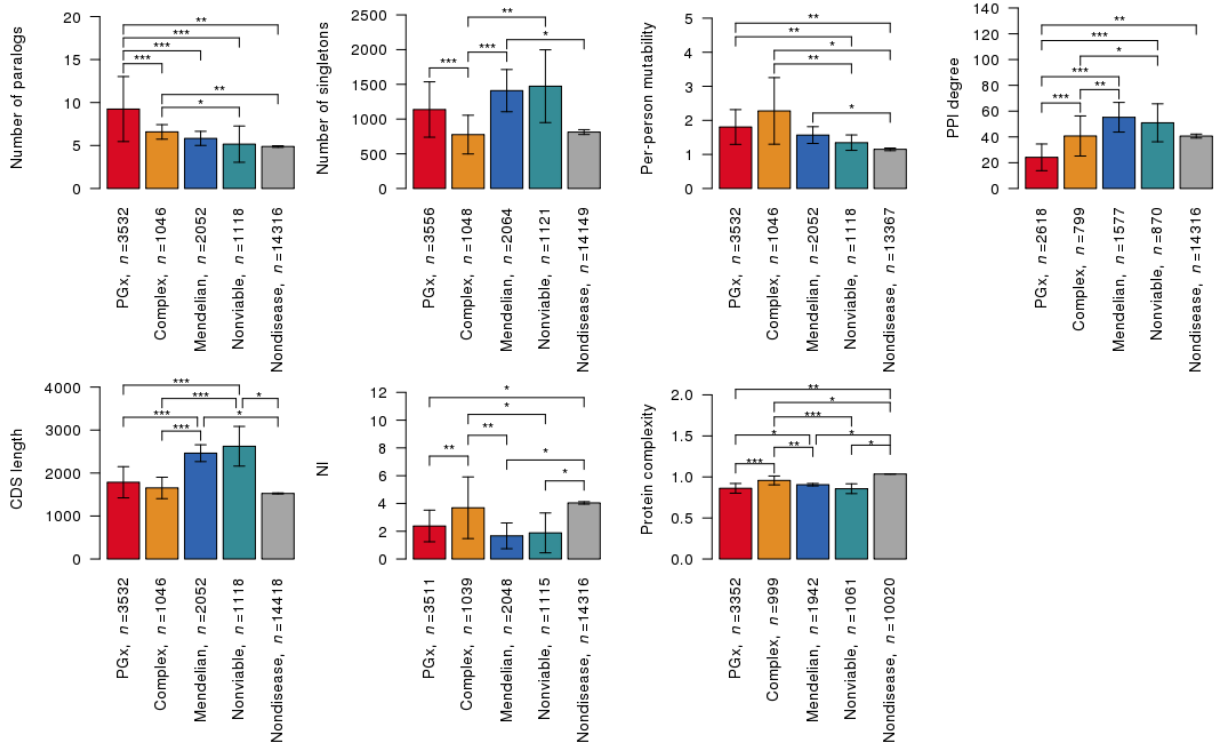


그림 10. 약물, 복합질환, 희귀질환, 생존 불가능, 그리고 비질환 유전자에 대한 일곱가지 유전적 분자 특성의 특성화. 하위 범주의 평균 유전자 특징에 대한 분포를 표현함. Wilcoxon test $*p < 0.1$, $**p < 0.05$, $***p < 0.01$.

3.3.4 약물, 복합질환, 및 희귀질환 유전자의 유전적 조성

그림 11 는 서로 다른 카테고리 간 분자적 특성을 바탕으로 한 유전적 조성 (genomic landscape)을 나타낸다. 일곱 개의 서로 다른 분자적 특성 중, 약물 – 복합 질환 (그림 11A), 약물 – 희귀 질환 (그림 11B), 그리고 복합 – 희귀 질환 (그림 11C) 을 구분하는 데 가장 큰 기여를 하는 세 가지 분자적 특성을 골라 그림으로 나타내었다. 복합 질환 유전자는 적은 수의 singleton 갯수 (x 축)와 paralog 갯수 (y 축), 그리고 높은 단백질 복잡도 (점 크기)를 가지는 반면, 약물 유전자는 많은 수의 paralog 갯수와 보통의 singleton 갯수, 단백질 복잡도를 가지고 있다. 이는 복합 질환 유전자에는 약물 유전자 대비 환경적으로 잘 적응한 변이가 무질서한 영역에서 상대적으로 높은 선택 압력 하에 위치하는 특성을 가진다는 것을 의미한다. 약물 유전자와 희귀 질환 유전자를 가장 잘 구분할 수 있는 특성으로는, 희귀 질환 유전자에서 상대적으로 긴 CDS 길이, 높은 단백질 상호 작용 정도, 그리고 적은 paralog 의 갯수가 있었다 (그림 11B). 예상대로, 복합 질환 유전자는 희귀 질환 유전자 대비 낮은 단백질 상호 작용 정도, 짧은 CDS 길이, 그리고 적은 수의 singleton 갯수를 보였다 (그림 11C).

일곱 가지 분자적 특성 중 CDS 의 길이를 고려했을 때 희귀

질환 유전자에 대한 예측 성능이 가장 많이 증가했고, 반면 paralog 의 갯수로 보정했을 때 약물, 그리고 복합 질환의 유전자에 대한 예측 성능이 평균적으로 가장 많이 증가하였다. 복합 질환의 경우, 하위 카테고리의 범주가 너무 상이하여 하나의 분자적 특성으로 구분하기 어려웠는데, 세부적으로는 복합 질환 내 정신 질환 관련 유전자 카테고리나 발달 질환 관련 유전자 카테고리에서 높은 수의 singleton 갯수와 긴 CDS 길이의 특성을 보였고, 이는 희귀 질환 카테고리의 특성과 매우 유사했다. 이러한 결과는 발달 질환과 주요 정신 질환의 높은 유전력이 멘델리안 유전자에서 예상되는 패턴과 매우 유사하다는 기존 보고 (해당 질환들의 복잡도 높은 특성상 반드시 멘델의 유전 법칙을 따르는 것은 아니지만)와 일치하는 경향성을 갖는다. 비슷하게, 약물 유전자 중 타겟, 그리고 효능 관련 유전자의 카테고리에서 적은 수의 paralog 갯수와 많은 수의 singleton 갯수를 보여, 희귀 질환 유전자와 매우 비슷한 특성을 보였다. 복합 질환 내 약물 유전체 관련 하위 카테고리는 예상대로 약물 유전자 카테고리와 매우 유사한 분자적 특성을 보였다. 약물 유전자 중 carrier 카테고리는 매우 독특한 특성의 분자적 특성을 보였는데, 적은 수의 paralog 와 singleton, 짧은 CDS 길이, 높은 단백질 복잡도의 특성을 나타냈다.

개인 별 변이성의 개념을 추가하여 그룹 별 유전적 조성을 4

차원 그래프로 나타내보면, 약물, 복합질환 및 희귀질환 카테고리에 속하는 유전자 카테고리가 서로 다른 영역으로 잘 구분되는 것을 확인할 수 있다 (그림 11D).

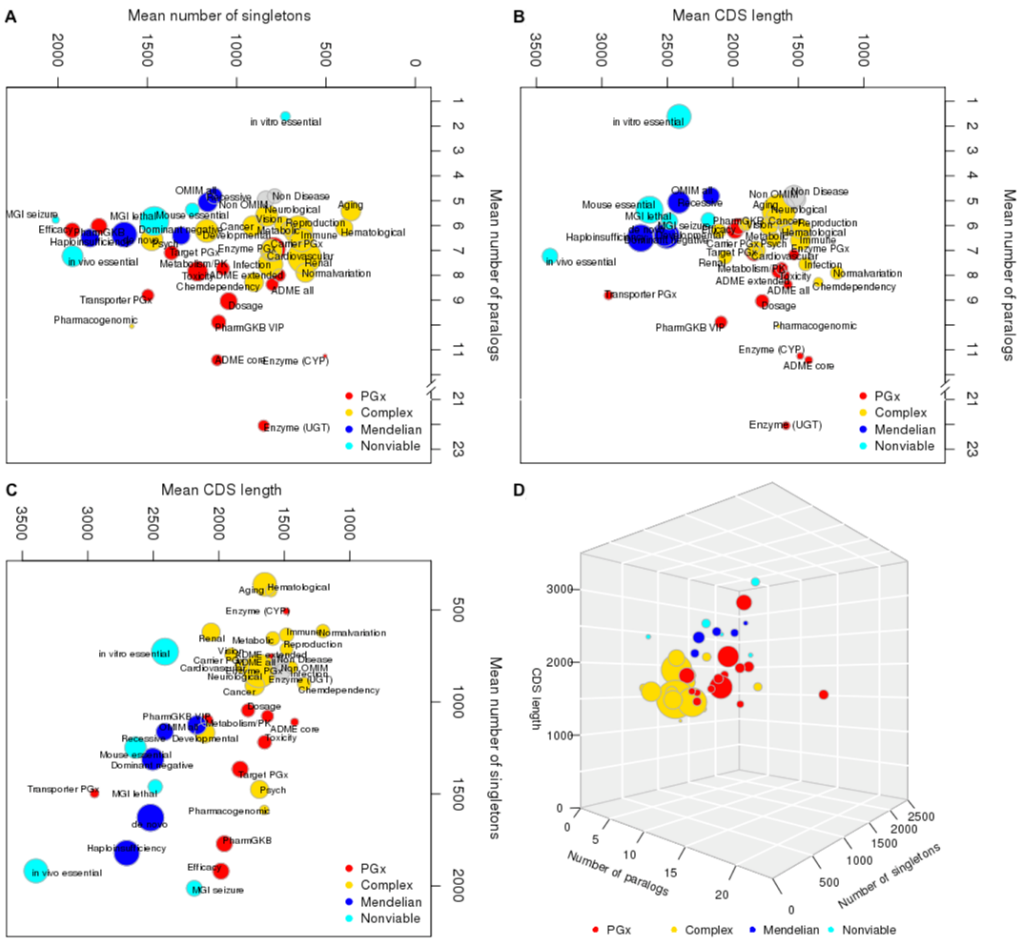
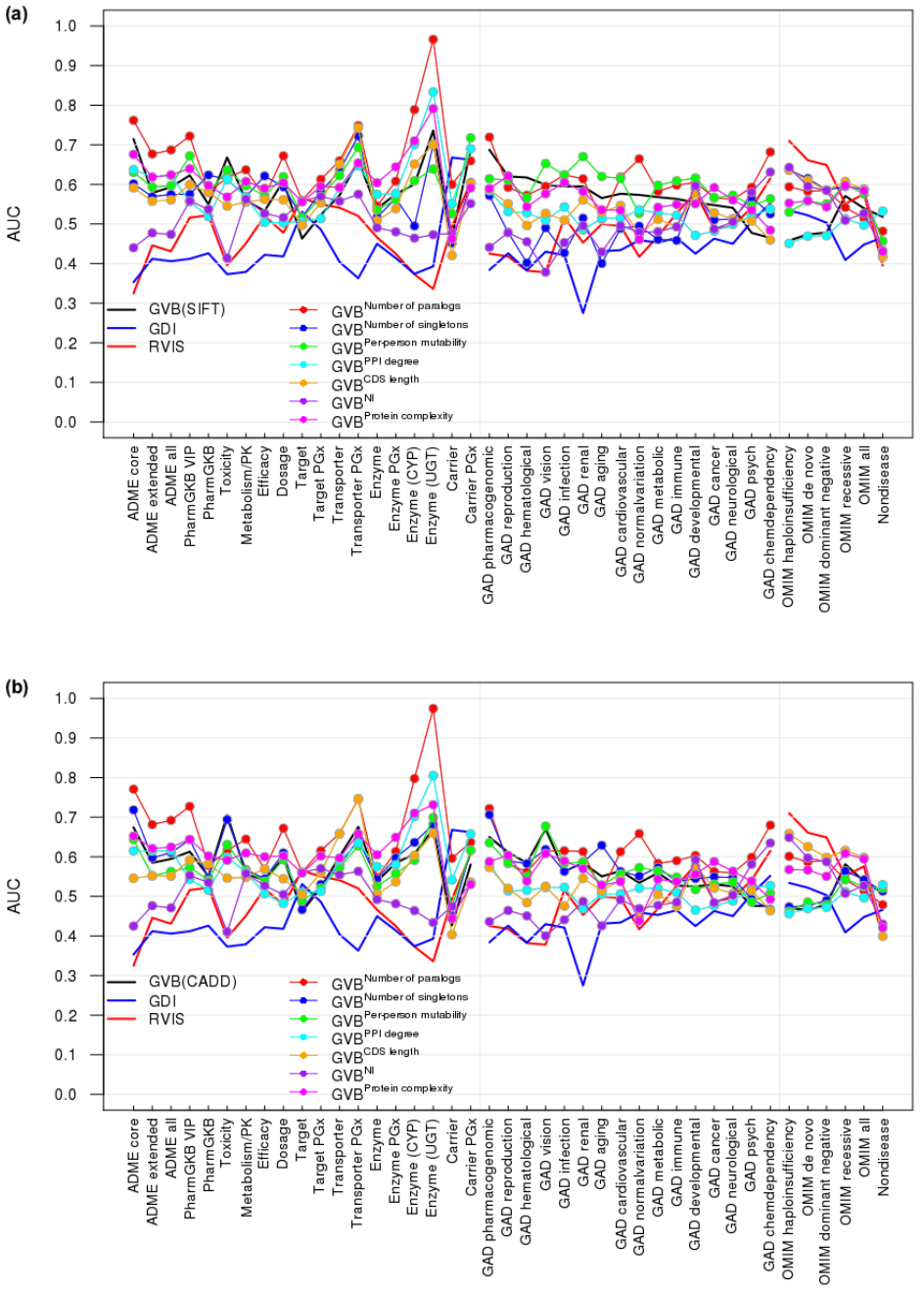


그림 11. 약물, 복합질환, 희귀질환의 하위 범주 간 유전적 분자 특성의 유전적 조성. 유전적 분자 특성의 유전적 조성은 (A) 약물 vs 복합질환 유전자, (B) 약물 vs 희귀질환 유전자, 그리고 (C) 복합 vs 희귀질환 유전자를 분류함. 평균 singletons 갯수 (유전적 다양성에 대한 증거) 와 paralogs 의 갯수 (선택적 압력에 대한 증거) 그리고 평균 CDS 길이가 수직 및 수평축에, 그리고 단백질 상호작용 정도와 단백질 복잡도는 원의 크기로 표현되어 있다. (D) 개인 별 변이 순응도를 포함한 네 가지 유전적 특성을 사용했을 때의 표현형 별 유전적 조성을 시각화 함. PGx, pharmacogenetic; CDS, coding sequence.

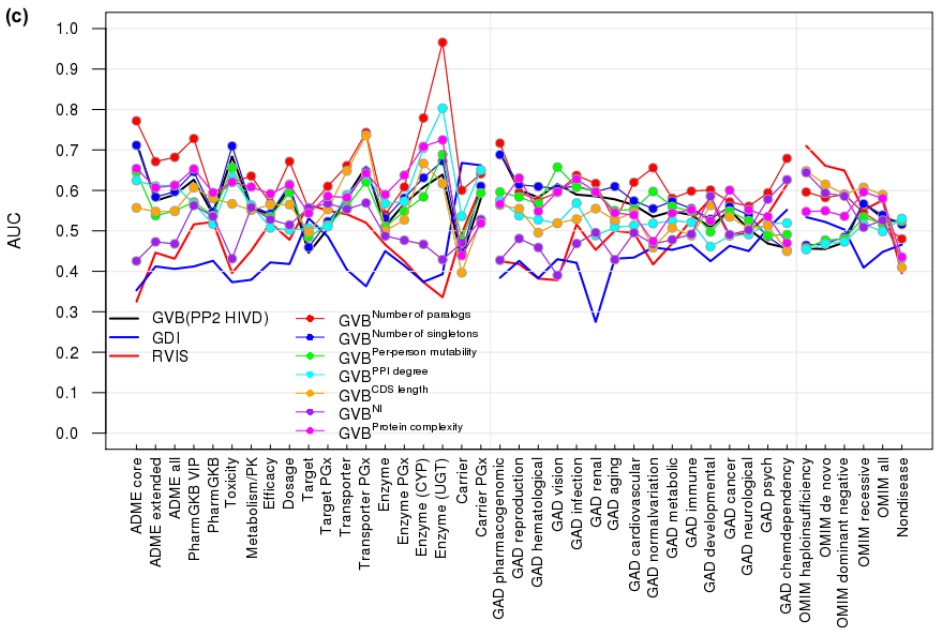
3.4 고찰

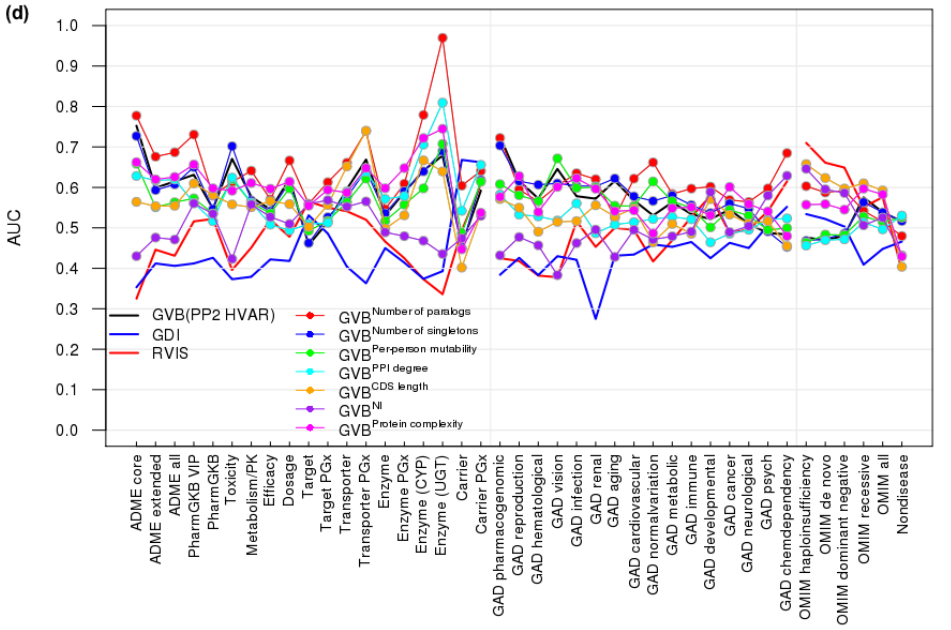
본 연구에서는 다양한 약물 및 질병 유전자 카테고리 내에서 유전자 수준 부담 점수의 실전적 유용성을 평가하였다. 현재 보고된 주요한 유전자 단위 점수들와 비교했을 때, GVB 는 특히 개인 간, 그리고 인종 간 유전적 다양성이 큰 약물 유전체 분야에 높은 활용도를 보였다. 이 밖에도 복합 질환 유전자의 예측에 상대적으로 높은 성능을 보였는데, 이는 만성 질환이 유전적 요인 뿐만 아니라 외부 환경으로부터의 자극과 생활 방식 등 다양한 요인의 복합적인 작용에 의해 발생한다는 측면에서 희귀 질환 보다는 약물 유발 부작용과 비슷하기 때문일 수 있을 것으로 추측된다.

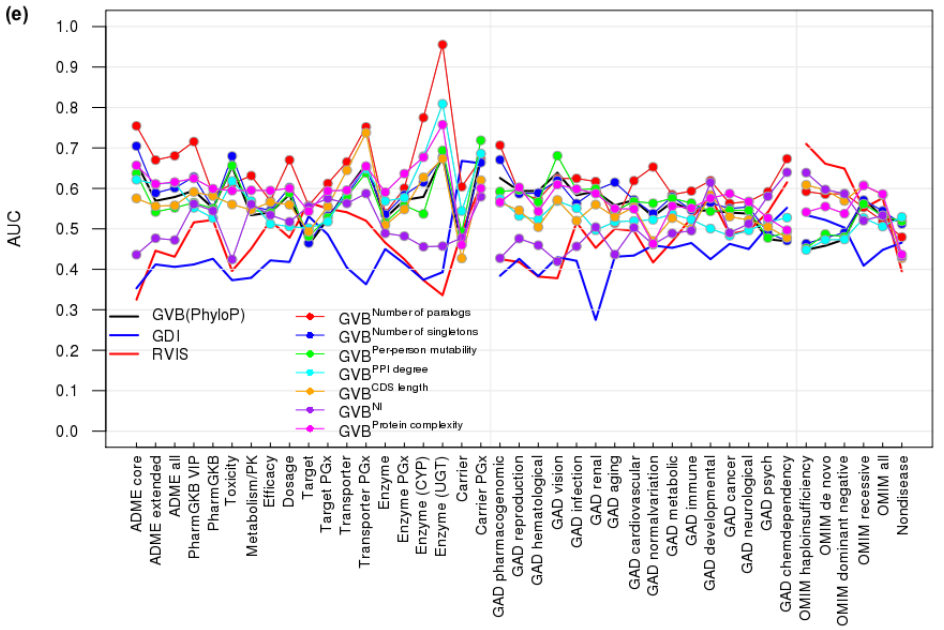
본 연구에서의 GVB 는 SIFT 알고리듬을 기반으로 한 결과를 서술하였지만, 다양한 *in silico* 예측 알고리듬을 기반으로 계산한 GVB 점수를 바탕으로 동일한 분석을 수행하였을 때, SIFT 알고리듬 기반의 분석 결과와 매우 유사한 예측 성능과 트렌드가 반복되는 것을 확인할 수 있다 (그림 12). 이는 변이 점수의 종류에 구애받지 않는다는 측면에서 GVB 점수로 도출한 결과에 대한 안정성을 반증한다.

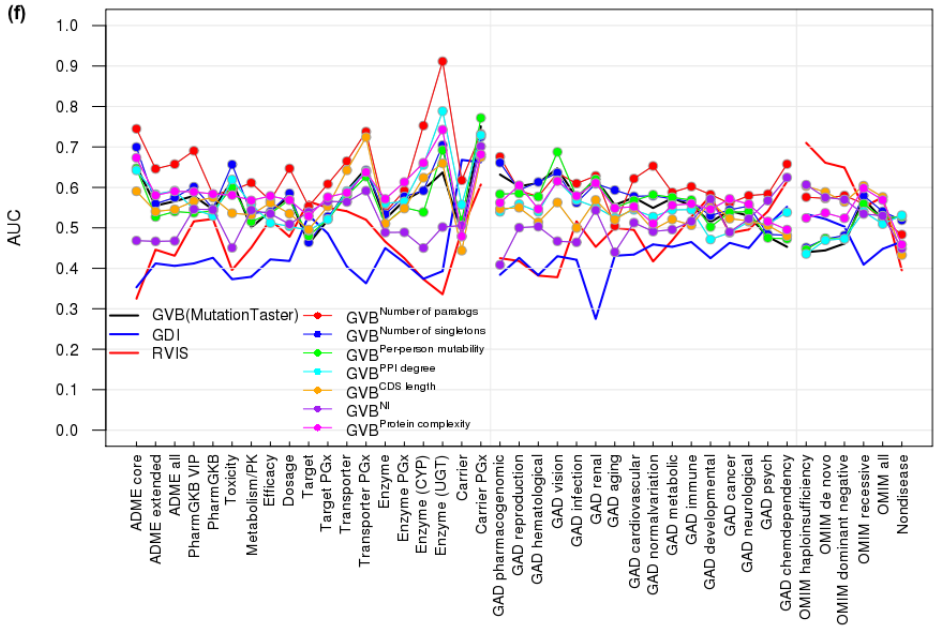

(a)
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GVB(SIFT)
GDI
RVIS
GVB Number of paralogs
GVB Number of singletons
GVB Per-person mutability
GVB PPI degree
GVB CDS length
GVB NI
GVB Protein complexity
ADME core
ADME extended
ADME all
PharmGKB VIP
PharmGKB
Toxicity
Metabolism/PK
Efficacy
Dosage
Target
Target PGx
Transporter
Transporter PGx
Enzyme
Enzyme PGx
Enzyme (CYP)
Enzyme (UGT)
Carrier
Carrier PGx
GAD pharmacogenomic
GAD reproduction
GAD hematological
GAD vision
GAD infection
GAD renal
GAD aging
GAD cardiovascular
GAD normalvariation
GAD metabolic
GAD immune
GAD developmental
GAD cancer
GAD neurological
GAD psych
GAD chemdependency
OMIM haploinsufficiency
OMIM de novo
OMIM dominant negative
OMIM recessive
OMIM all
Nondisease

(b)
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GVB(CADD)
GDI
RVIS
GVB Number of paralogs
GVB Number of singletons
GVB Per-person mutability
GVB PPI degree
GVB CDS length
GVB NI
GVB Protein complexity
ADME core
ADME extended
ADME all
PharmGKB VIP
PharmGKB
Toxicity
Metabolism/PK
Efficacy
Dosage
Target
Target PGx
Transporter
Transporter PGx
Enzyme
Enzyme PGx
Enzyme (CYP)
Enzyme (UGT)
Carrier
Carrier PGx
GAD pharmacogenomic
GAD reproduction
GAD hematological
GAD vision
GAD infection
GAD renal
GAD aging
GAD cardiovascular
GAD normalvariation
GAD metabolic
GAD immune
GAD developmental
GAD cancer
GAD neurological
GAD psych
GAD chemdependency
OMIM haploinsufficiency
OMIM de novo
OMIM dominant negative
OMIM recessive
OMIM all
Nondisease


(c)
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GVB(PP2 HIVD)
GDI
RVIS
GVB Number of paralogs
GVB Number of singletons
GVB Per-person mutability
GVB PPI degree
GVB CDS length
GVB NI
GVB Protein complexity
ADME core
ADME extended
ADME all
PharmGKB VIP
PharmGKB
Toxicity
Metabolism/PK
Efficacy
Dosage
Target
Target PGx
Transporter
Transporter PGx
Enzyme
Enzyme PGx
Enzyme (CYP)
Enzyme (UGT)
Carrier
Carrier PGx
GAD pharmacogenomic
GAD reproduction
GAD hematological
GAD vision
GAD infection
GAD renal
GAD aging
GAD cardiovascular
GAD normalvariation
GAD metabolic
GAD immune
GAD developmental
GAD cancer
GAD neurological
GAD psych
GAD chemdependency
OMIM haploinsufficiency
OMIM de novo
OMIM dominant negative
OMIM recessive
OMIM all
Nondisease


(d)
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GVB(PP2 HVAR)
GDI
RVIS
GVB Number of paralogs
GVB Number of singletons
GVB Per-person mutability
GVB PPI degree
GVB CDS length
GVB NI
GVB Protein complexity
ADME core
ADME extended
ADME all
PharmGKB VIP
PharmGKB
Toxicity
Metabolism/PK
Efficacy
Dosage
Target
Target PGx
Transporter
Transporter PGx
Enzyme
Enzyme PGx
Enzyme (CYP)
Enzyme (UGT)
Carrier
Carrier PGx
GAD pharmacogenomic
GAD reproduction
GAD hematological
GAD vision
GAD infection
GAD renal
GAD aging
GAD cardiovascular
GAD normalvariation
GAD metabolic
GAD immune
GAD developmental
GAD cancer
GAD neurological
GAD psych
GAD chemdependency
OMIM haploinsufficiency
OMIM de novo
OMIM dominant negative
OMIM recessive
OMIM all
Nondisease


(e)
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GVB(PhyloP)
GDI
RVIS
GVB Number of paralogs
GVB Number of singletons
GVB Per-person mutability
GVB PPI degree
GVB CDS length
GVB NI
GVB Protein complexity
ADME core
ADME extended
ADME all
PharmGKB VIP
PharmGKB
Toxicity
Metabolism/PK
Efficacy
Dosage
Target
Target PGx
Transporter
Transporter PGx
Enzyme
Enzyme PGx
Enzyme (CYP)
Enzyme (UGT)
Carrier
Carrier PGx
GAD pharmacogenomic
GAD reproduction
GAD hematological
GAD vision
GAD infection
GAD renal
GAD aging
GAD cardiovascular
GAD normalvariation
GAD metabolic
GAD immune
GAD developmental
GAD cancer
GAD neurological
GAD psych
GAD chemdependency
OMIM haploinsufficiency
OMIM de novo
OMIM dominant negative
OMIM recessive
OMIM all
Nondisease


(f)
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
GVB(MutationTaster)
GDI
RVIS
GVB Number of paralogs
GVB Number of singletons
GVB Per-person mutability
GVB PPI degree
GVB CDS length
GVB NI
GVB Protein complexity
ADME core
ADME extended
ADME all
PharmGKB VIP
PharmGKB
Toxicity
Metabolism/PK
Efficacy
Dosage
Target
Target PGx
Transporter
Transporter PGx
Enzyme
Enzyme PGx
Enzyme (CYP)
Enzyme (UGT)
Carrier
Carrier PGx
GAD pharmacogenomic
GAD reproduction
GAD hematological
GAD vision
GAD infection
GAD renal
GAD aging
GAD cardiovascular
GAD normalvariation
GAD metabolic
GAD immune
GAD developmental
GAD cancer
GAD neurological
GAD psych
GAD chemdependency
OMIM haploinsufficiency
OMIM de novo
OMIM dominant negative
OMIM recessive
OMIM all
Nondisease

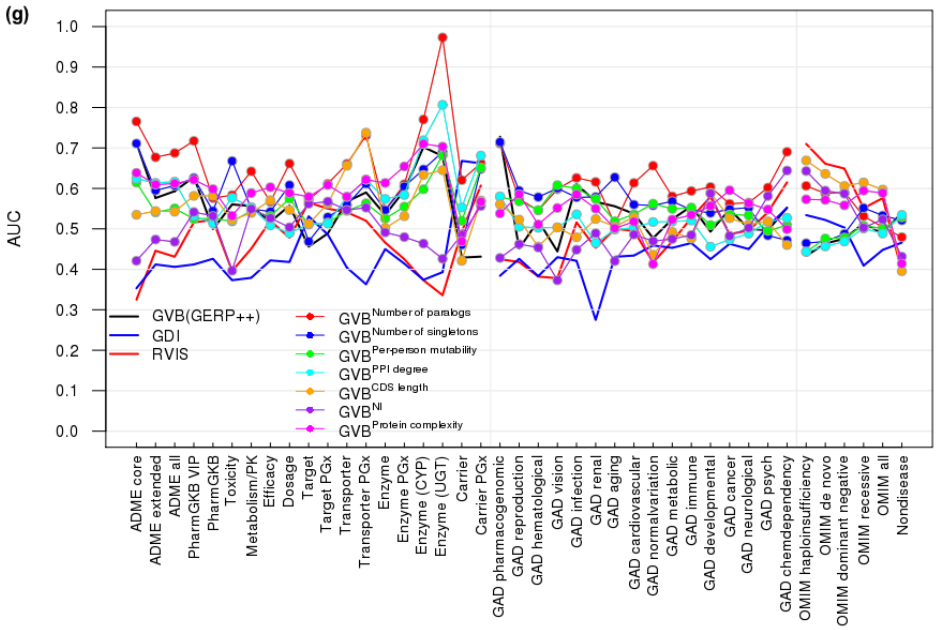


그림 12. 일곱가지 *in silico* 예측 방법을 사용하여 계산된 GVB 의 예측 성능 평가. GVB 는 (*a*) SIFT, (*b*) CADD, (*c*) PolyPhen2 HIVD, (*d*) PolyPhen2 HVAR, (*e*) PhyloP, (*f*) MutationTaster, 그리고 (*g*) GERP 를 사용하여 계산되었으며, 일곱가지 유전적 분자 특성에 의해 보정되었다.: paralog 의 갯수, singleton 의 갯수, 개인 별 변이 순응도, 단백질 상호작용 정도, CDS 길이, McDonald–Kreitman 중립 지수(NI), 그리고 단백질 복잡도. AUC: area under the receiver operating characteristic curve; GAD, Genetic Association Database; ADME, Absorption, Distribution, Metabolism, and Excretion; PK, pharmacokinetic; PGx, pharmacogenetic; OMIM, Online Mendelian Inheritance in Man.

GVB 가 사람 별로 점수를 제공한다는 이점을 살려서, paralog 의 갯수가 많을 수록 진화적 제약 (evolutionary constraints)은 약할 것이라는 일반적인 가정에 따라 [81], 한 유전자에 대한 유전적 다양성이 높을 수록 (GVB 점수의 분산이 클 수록) paralog 의 갯수도 증가하는 경향성이 있는 지 확인해보았다. 결론적으로는, 2504 명에서 GVB 의 차이가 클 수록 paralog 의 갯수가 증가하는 양의 상관 관계를 확인할 수 있었는데 (그림 13, Kendall's tau=0.085, $p<0.0001$), 이는 GVB 가 진화적 유전 변이의 다양성에 대한 정량적 척도로 사용될 수 있음을 시사한다.

유전자 카테고리 중 생존할 수 없는 (nonviable) 유전자는 거의 대부분 희귀 질환에서 보이는 특성과 비슷하거나 혹은 더 극도로 치우친 트렌드를 보이는 경향성이 있었다. 흥미롭게도 비질환성 (non-disease) 유전자에서 예상과 달리 약물 및 질환 유전자 보다 낮은 paralog 의 갯수와 높은 단백질 상호작용 정도를 보이는 경향성이 있었는데, 이는 비질환의 카테고리에 살아남지 못한 (lethal) 훨씬 더 강력한 선택적 압력를 받는 허브 (hub) 유전자가 포함되어 있음을 시사한다 [82].

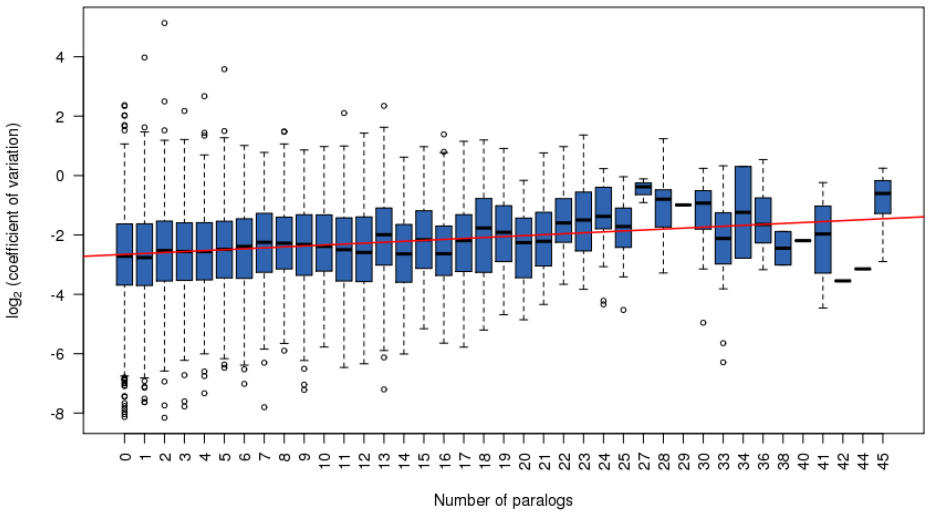


그림 13. 1000 지놈 프로젝트의 2504 명에서 계산된 샘플 간 GVB 점수의 편차 와 paralog 갯수 간 상관 관계. Kendall τ = 0.085 (regression p = 3.32e−56).

오랜 시간 동안 서로 다른 유전적 배경을 가지는 유전자들은 서로 다른 유전적 특성을 이어 받을 것 이라고 가정되어 왔다. 본 연구는 약물, 복합 질환, 희귀 질환 간 직관적으로 추정되어온 유전적 차이에 대해서 체계적으로 탐색한 첫번째 연구이다. 서로 다른 카테고리의 표현형은 매우 상이한 유전적 특성을 가지고 있기 때문에 목적에 따라 질환 특이적인 유전자의 분자적 특성을 확인하고 해당 질환에 적합한 접근법을 찾아 최적화된 방법을 반영하는 것이 중요하다. 본 연구는 다양한 유전적 시나리오 하에서 다양한 유형의 유전적 특징을 기반으로 정교한 체계가 적용되어야 한다는 점에서 한계가 있다. 유전자 단위 뿐 아니라 엑손, 도메인 등의 세분화 된 영역을 기준으로 한 평가가 추가로 이루어져야 한다. 더불어 연구 결과의 보다 정확한 임상적 해석을 돕기 위해서는 copy number variation 등 대규모 DNA 변화에 대한 효과도 포함되어야 한다.

4 *NUDT15* 과 *TPMT* 에 모두 변이를 가지고 있지 않은 소아 백혈병 환자에서 치오퓨린 연관 유전자의 탐색

4.1 연구배경

NUDT15 과 *TPMT* 는 소아아 급성림프구성 백혈병 환자에서 6-mercaptopurine (6-MP) 약물 유발 부작용과의 연관성이 가장 잘 정립되어있는 유전자이다. 유럽 인종의 경우, 호중구 감소증 (neutropenia) 또는 백혈구 감소증 (leukopenia) 등 심각한 치오퓨린 유발 부작용의 약 50%가 해당 두 유전자에서 발견되는 변이로 설명된다고 보고되어 있다 [83]. 이에 따라, CPIC [84]은 해당 두 유전자에서 발견되는 변이들을 기반으로 실제 임상에서 약물유전체 테스트의 실질적인 이행이 가능하도록하는 증거 기반 가이드라인 (evidence-based guideline)을 제공하고 있다 [32, 33].

현재, 임상에서 6-MP 의 용량은 *NUDT15* 과 *TPMT* 에서 발견된 알려진 위험 변이 (risk variant)들에 기반하여 조절되고 있다. 그러나, 해당 두 유전자에 변이를 가지고 있지 않은 소아 밸혈병 환자의 상당수가 여전히 생명을 위협할만큼 심각한 부작용을 겪고 있고, 이에 따라 약물 용량을 줄이거나 혹은 약물 복용을 중단하면서 치료 실패와 재발을 겪고 있다. 따라서, 소아 백혈병 환자의 치료를 돕기 위해서는

NUDT15 과 *TPMT* 이외에 6-MP 약물 독성과 연관된 새로운 유전 변이에 대한 탐색이 매우 필요한 실정이다.

본 연구에서는 엑솜 시퀀싱 데이터를 이용하여 *NUDT15* 과 *TPMT* 에 변이를 가지고 있지 않은 소아 백혈병 환자에서 6-MP 약물 부작용과 연관된 새로운 유전 변이를 탐색하는 것을 목표로 한다. 임상적으로 혈액 독성 (hematological toxicity)에 대한 주요 지표로 활용할 수 있는 유지요법 중 마지막 사이클의 예측대비 실제 6-MP 약물 투약 용량 (DIP)을 활용하여 약물 독성과 연관된 새로운 후보 유전자자를 체계적으로 탐색하였다.

4.2 재료 및 방법론

4.2.1 환자군

유지 요법 동안 6-MP 약물을 투여 받은 320 명의 한국인 소아 백혈병 환자를 모집하였다. 320명 중 2018년 2월 이전에 시퀀싱된 244명 (발견 코호트)은 2 개 병원에서 (서울대 병원 [SNUH]과 아산 의료원 [AMC]), 2018 년 10 월 이후부터 2019 년 11 월까지 시퀀싱된 76 명 (복제 코호트)은 3개 병원에서 (서울대병원, 아산 의료원, 삼성 의료원 [SMC]) 모집되었다. 모든 환자는 제외 사유(*i.e.*, relapse of the disease, stem cell transplantation, Burkitt's lymphoma, mixed phenotype acute leukemia, infant ALL, 또는 very high risk of ALL)가 있는 경우 제외되었다. 약물 독성은 유지요법 중 마지막 사이클에서의 DIP 를 기반으로 추정되었다. 동아시아인의 경우 다른 인종에 비해 매우 낮은 6-MP 약물 용량이 요구되므로 [85], 예측 용량 대비 실제 투약 용량이 25% 이하인 경우 6-MP 약물에 민감한 위험군으로 분류되었다 [86]. 이전 연구에서 환자군에 대한 자세한 설명과 DIP 측정 방법을 기술하였다 [16]. 본 연구는 서울대 병원, 아산 의료원, 삼성 의료원의 IRB 승인 절차를 거쳤으며, 모든 환자에게 서면 동의를 받았다.

4.2.2 엑솜 시퀀싱과 데이터 분석

320 명의 소아 백혈병 환자에 대한 엑솜 시퀀싱과 생물정보학 분석 과정은 이전 연구에서 자세히 기술되었다 [16]. *NUDT15* 에서 발견된 스타 대립 유전자가 등록되지 않은 두 개의 변이 (rs780144127 와 13: 48611982 A>G)는 Sanger sequencing 을 통해 확인되었고, 그 중 위양성 변이는 추후 분석에서 제외되었다. 2019 년 2 월에 업데이트된 CPIC 가이드라인에 따라 *NUDT15**9 의 기능은 '불분명한 (uncertain)' 에서 '기능이 저하된 (no function)'으로 변경되었으며 [83], 이에 따라 한 명의 환자가 'poor metabolizer'로 재분류되었다. 현 분석은 *NUDT15* 과 *TPMT* 모두에서 변이가 발견되지 않은 240 명의 normal metabolizer (188 명의 발견, 52 명의 복제 코호트)를 대상으로 수행되었다. 발견 단계에서는 SnpEFF (http://snpeff.sourceforge.net) [87]를 사용하여 변이의 기능을 예측하였으며, 이 중 아미노산 변경에 강력한 영향을 미칠 것으로 예상되는 변이 (missense, nonsense, splice-site, frameshift, and in-frame insertion and deletion variants)만이 선택되었다 (그림 14).

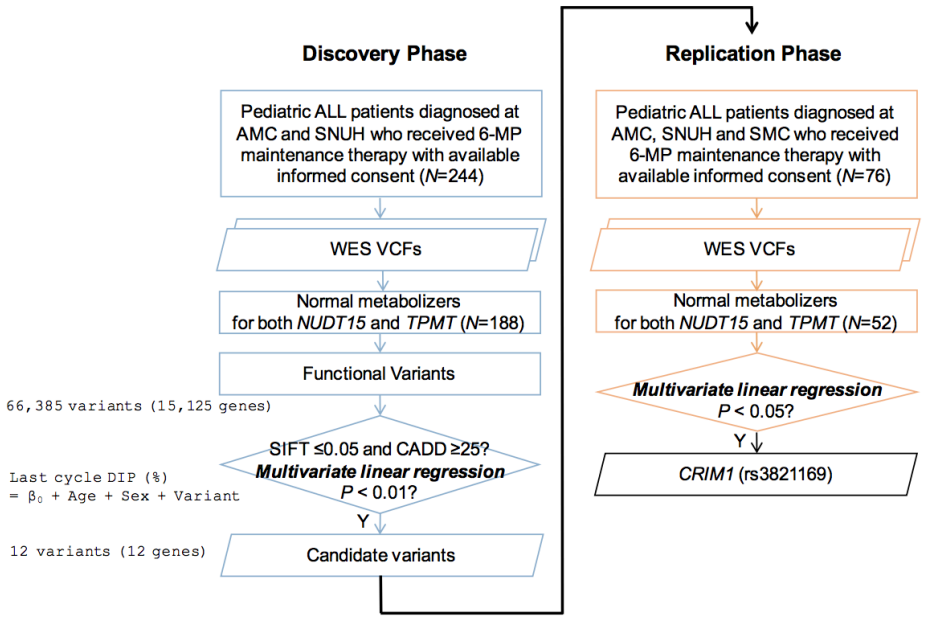


그림 14. 발견 및 복제 단계 데이터 분석의 요약도. ALL, acute lymphoblastic leukemia; WES, whole-exome sequencing; SIFT, sorting intolerant from tolerant; CADD, combined annotation-dependent depletion; VCF, variant call format; DIP, dose intensity percentage; NM, normal metabolizer; AMC, Asan Medical Center; SNUH, Seoul National University Hospital; SMC, Samsung Medical Center; 6-MP, 6-mercaptopurine.

DIP 모델의 나이와 성별을 보정한 다중 회귀 분석을 통하여 159 개의 유전자에 위치한 185 개의 변이 (p<0.01) 중, 두 개의 *in silico* 예측 방법에서 유해하다고 예측된 (*i.e.*, SIFT score [59] ≤0.05 그리고 CADD score [42] ≥25) 변이를 별도의 복제 코호트에서 다중 회귀 분석을 사용하여 재평가하였다. 결과적으로 additive 와 recessive 모델 모두에서 유의한 결과를 보여 같은 결과가 재현된 한개 변이 (Cysteine-Rich Transmembrane BMP Regulator 1 (*CRIM1*) 유전자에서 발견된 rs3821169)를 최종 후보 변이로 확인하였다. 최종 후보 변이에 대해서는 별도의 genotyping assay (SNPtype; Fluidigm, San Francisco, CA)를 수행하여 시퀀싱 결과에 오류가 없는지 재확인하였다. SNPtype assay 는 시퀀싱 후 추가로 blood DNA 확보가 가능했던 118 명에 대하여 수행하였다.

4.2.3 단일- 그리고 다중 유전자를 사용한 치오퓨린 독성 예측 정확도

흔한 변이와 드문 변이 모두의 통합 효과를 평가하기 위하여 유전자 수준 변이 점수 분석을 수행하였다. 모든 코딩 유전자에 대하여 각 유전자에서 발견된 코딩 변이의 SIFT 점수의 기하 평균을 계산하여 개인별 GVB 점수를 산출하였으며, 이 때 GVB^{G} 는 유전자 G 에 대한 GVB 점수를 의미한다. 6-MP DIP 를 예측하는 GVB^{NUDT15}, GVB^{TPMT}, 그리고

GVB_{CRIM1} 의 성능은 ROC 분석을 통해 체계적으로 평가되었으며, 발견, 복제, 그리고 혼합 코호트 각각에서의 AUC 값을 일곱 개의 DIP cutoffs (*i.e.*, 15%, 25%, 35%, 45%, 60%, 80%, 그리고 100%)에서 측정하였다. 단일 유전자의 효과는 다른 두 유전자에 변이를 가지는 경우에 미칠 수 있는 효과를 통제한 뒤 평가되었다. 다중 유전자 효과는 GVB^{A} 와 GVB^{B} 의 기하평균으로 정의된 $GVB^{A,B}$ 를 사용하여 체계적으로 평가되었다.

모든 통계 분석은 R 통계 패키지 (3.5.1 버전)를 사용하여 수행되었다. 특히 recessive model 에서의 *CRIM1* 변이 효과를 정확하게 평가하기 위하여 heterozygous rs3821169 의 효과를 무시한 GVB^{CRIM1} 가 계산되었다.

4.3 결과

4.3.1 환자군에 대한 설명

320 명의 소아 백혈병 환자 중 *NUDT15* 과 *TPMT* 모두에서 CPIC 에서 보고한 유해한 변이를 하나도 보유하지 않은 240 명 (발견 코호트 244 명 중 188 명과 복제 코호트 76 명 중 52 명)이 선택되었다. 표 7 은 *NUDT15* 과 *TPMT* 모두 WT 인 240 명에 대한 임상적 특성을 나타낸 표이다. 양쪽 유전자 모두 WT 이 아닌 군 (80 명)과 비교했을 때, 양쪽

모두 WT 인 군 (240 명)은 발견 코호트 [68.44±27.6 vs. 54.14±29.9 (mean±SD), *t*-test p=0.002]와 복제 코호트 [59.99±38.2 vs. 33.36±28.7, *t*-test p=0.001], 그리고 두 군을 모두 합친 군 모두에서 상대적으로 유의하게 높은 DIP 수치를 보였다. 이러한 결과는 소아 백혈병에서 치오퓨린 독성에 영향을 미친다고 알려진 두 개의 약물유전자 (*NUDT15* 과 *TPMT*)의 효과가 다시 한번 확인된 것임을 의미한다.

그러나 표 11 은 또한 양쪽 유전자 모두에 변이를 갖지 않는 군에서 발견 코호트의 4.8% (188 명 중 9 명), 그리고 복제 코호트의 23.1% (52 명 중 12 명) 각각이 치오퓨린 중증 독성 위험군 (DIP 〈 25%)으로, 그리고 63.8% (188 명 중 120 명)과 46.2% (52 명 중 24 명)이 치오퓨린 경증 독성 위험군 (DIP 〈 80%)으로 분류되었다는 것을 나타낸다. 이 때, 발견과 복제 코호트 에서 중증 독성 위험군의 비율에 차이가 있는 것은 단순히 복제 데이터가 부족했기 때문인 것으로 판단된다. 전체적으로 240 명 중 68.8% (165 명)에서 *NUDT15* 과 *TPMT* 만으로는 설명하지 못하는 치오퓨린에 대한 다양한 약물 반응성을 보인다는 것을 확인하였다.

표 11. *NUDT15* and *TPMT* 모두에 변이를 갖지 않는 소아 백혈병 환자에 대한 임상적 특성.

특성	발굴	복제	혼합
환자 수	188	52	240
나이, 년 †	6.9 ± 4.5	7.4 ± 4.5	7.0 ± 4.5
성별			
남	108	29	137
여	80	23	103
마지막 사이클 6-MP 용량 백분율, mg/m^2/day			
≤10	8.68 ± 1.5 (3)	6.29 ± 2.2 (5)	7.19 ± 2.2 (8)
>10 & ≤15	13.89 ± NA (1)	13.21 ± 1.8 (3)	13.38 ± 1.5 (4)
>15 & ≤25	18.52 ± 3.4 (5)	22.13 ± 1.3 (4)	20.12 ± 3.1 (9)
>25 & ≤35	29.95 ± 3.3 (16)	30.49 ± 0.8 (4)	30.06 ± 3.0 (20)
>35 & ≤45	39.54 ± 3.6 (8)	40.18 ± 2.6 (5)	39.79 ± 3.1 (13)
>45 & ≤60	52.71 ± 4.0 (41)	54.84 ± 2.8 (5)	52.94 ± 3.9 (46)
>60 & ≤80	70.79 ± 6.0 (55)	69.35 ± 5.1 (10)	70.57 ± 5.8 (65)
>80&≤100	90.87 ± 5.9 (35)	85.98 ± 5.0 (8)	89.96 ± 6.0 (43)*
>100	112.67 ± 16.6 (24)	122.66 ± 23.3 (8)	115.17 ± 18.7 (32)
Total	68.44 ± 27.6 (188)	59.99 ± 38.2 (52)	66.61 ± 30.3 (240)

†한 명에 대한 나이 정보는 얻을 수 없었음; 6-MP, 6-mercaptopurine; NA, not available; p 값은 t-tests 또는 χ^2 tests 중 적절한 것을 사용하여 계산됨. *$p<0.05$

4.3.2 *NUDT15*과 *TPMT* 이외의 치오퓨린 독성 후보 유전자

나이와 성별을 보정한 변이 수준의 다중 회귀 분석은 유전자 기능에 강력한 영향을 미칠 것으로 예측된 66,385 변이 (*i.e.*, 64,238 missense, 1,249 nonsense, 552 splice-site, 332 frameshift, 그리고 4 in-frame insertion and deletion)에 대해 수행되었다 (그림 14). 12 개 유전자에서 발견된 12 개의 후보 변이는 $p<0.01$ 의 cutoff 와 2 개의 *in silico* 예측 방법 (SIFT ≤ 0.05 그리고 CADD ≥ 25)을 적용하여 선택되었다. 사용된 샘플 숫자의 한계로 다중 검정을 통해 genome-wide 한 유의성을 얻을 수 없었기 때문에, 상대적으로 덜 엄격한 p cutoff 가 적용되었다.

표 12 는 치오퓨린 독성에 연관성을 보이는 12 개의 후보 유전자를 나타낸다. 이 중 *CRIM1* 유전자에서 발견된 rs3821169 변이만이 복제 코호트에서 additive (p=0.0483) 그리고 recessive (p=0.0132) 모델 모두에서 통계적으로 유의한 결과를 재현하였다 (표 13). 12 개 변이 중 10 개 변이의 경우 적은 샘플 수와 드물게 발견되는 특성 때문에 recessive 모델을 적용할 수 없었다.

표 12. 발견 단계에서 *NUDT15*과 *TPMT* 모두에 변이를 갖지 않는 188 명의 환자로부터 얻어진 12 개의 후보 변이에 대한 평가.

변이: 위험 인자	유전자 이름	SIFT 점수	CADD 점수	ExAC EAS 빈도	변이 보유자	6-MP 용량 백분율 (%)			additive 모델		recessive 모델	
						보유자	비보유자	ANOVA p	효과 크기	$p^{\dagger}$	효과 크기	$p^{\dagger}$
rs3821169:T	*CRIM1*	0	25.3	0.243	88	64.41±27.7	71.99±27.0	0.015	–9.07	0.0079	–21.09	0.0248
rs191083003:T	*FSIP2*	0.01	26.7	3.46E-03	3	25.79±21.1	69.13±27.1	0.007	–46.98	0.0033	NA	NA
rs67877771:G	*IQCG*	0.04	26.2	0.215	59	61.13±25.3	71.79±28.0	0.010	–10.68	0.0086	–22.55	0.2531
rs200125400:A	*SLC22A5*	0	32	2.39E-03	2	16.44±0.7	69.00±27.2	0.007	–52.63	0.0069	NA	NA
rs141145196:A	*TOP1MT*	0.03	27.1	4.76E-03	2	19.46±17.7	68.97±27.2	0.011	–54.90	0.0061	NA	NA
rs61758536:A	*SPAG8*	0	26	0.052	28	56.14±26.6	70.60±27.2	0.010	–14.74	0.0087	NA	NA
rs181036640:A	*DPP7*	0	28.7	0.011	4	31.90±28.4	69.24±27.1	0.007	–37.27	0.0071	NA	NA
rs34337292:C	*OR9Q2*	0	25.9	0.068	48	59.46±28.7	71.52±26.5	0.002	–11.98	0.0044	–37.44	0.0195
rs200982819:A	*SLC15A3*	0	29.7	0.028	7	42.40±18.8	69.45±27.4	0.005	–24.41	0.0056	–58.42	0.0340
rs144612495:T	*GOLGA3*	0.02	25.7	4.00E-03	2	18.16±6.0	68.98±27.2	0.009	–55.00	0.0049	NA	NA
rs12587478:T	*KLHL33*	0	25	0.059	11	44.20±27.5	69.95±26.9	0.005	–22.01	0.0034	–17.75	0.5218
rs746000108:T	*INSR*	0.01	25	5.01E-04	2	18.11±15.8	68.98±27.2	0.009	–50.77	0.0097	NA	NA

$^{\dagger}p$ 값은 다중 선형 회귀를 통해 얻어짐; SIFT, sorting intolerant from tolerant; CADD, combined annotation-dependent depletion; EAS, East Asians; ExAC, Exome Aggregation Consortium; NA, Not Available.

표 13. 발견 단계에서 얻어진 12 개의 후보 변이에 대한 복제 단계 (*N*=52)에서의 평가 결과.

						6-MP 용량 백분율 (%)			additive 모델		recessive 모델	
변이: 위험 인자	유전자 이름	SIFT 점수	CADD 점수	ExAC EAS 빈도	변이 보유자	보유자	비보유자	ANOVA *p*	보유자	비보유자	ANOVA *p*	보유자
rs3821169:T	*CRIM1*	0	25.3	0.243	25	55.14±40.5	64.47±36.1	0.118	- 16.55	**0.0483**	−52.27	**0.0132**
rs191083003:T	*FSIP2*	0.01	26.7	3.46E-03	1	8.82±NA	60.99±37.8	0.178	−39.52	0.2877	NA	NA
rs67877771:G	*IQCG*	0.04	26.2	0.215	17	65.98±24.5	57.07±43.3	0.676	0.52	0.9492	−16.54	0.3899
rs200125400:A	*SLC22A5*	0	32	2.39E-03	2	56.21±37.1	60.14±38.6	0.888	14.29	0.5975	NA	NA
rs141145196:A	*TOP1MT*	0.03	27.1	4.76E-03	1	142.05± NA	58.38±36.7	**0.028**	68.60	0.0626	NA	NA
rs61758536:A	*SPAG8*	0	26	0.052	5	42.83±37.6	61.81±38.2	0.295	−17.38	0.3186	NA	NA
rs181036640:A	*DPP7*	0	28.7	0.011	1	88.24± NA	59.43±38.3	0.460	41.32	0.2688	NA	NA
rs34337292:C	*OR9Q2*	0	25.9	0.068	14	53.67±28.9	62.31±41.2	0.474	−5.15	0.6550	NA	NA
rs200982819:A	*SLC15A3*	0	29.7	0.028	1	80.13± NA	59.59±38.4	0.599	11.78	0.7521	NA	NA
rs144612495:T	*GOLGA3*	0.02	25.7	4.00E-03	1	88.24± NA	59.43±38.3	0.460	41.32	0.2688	NA	NA
rs12587478:T	*KLHL33*	0	25	0.059	2	30.82±1.2	61.15±38.5	0.275	−20.96	0.4436	NA	NA
rs746000108:T	*INSR*	0.01	25	5.01E-04	0	NA	NA	NA	NA	NA	NA	NA

†*p* 값은 다중 선형 회귀를 통해 얻어짐; SIFT, sorting intolerant from tolerant; CADD, combined annotation-dependent depletion; EAS, East Asians; ExAC, Exome Aggregation Consortium; NA, Not Available.

4.3.3 *CRIM1* 변이와 치오퓨린 독성간 연관성 평가

CRIM1 rs3821169 변이 보유자는 발견 (p=0.007), 복제 (p=0.048), 그리고 혼합군 (p<0.001) 모두에서 유의하게 낮은 6-MP DIP 수치를 보였다 (그림 15). 이러한 강력한 연관은 recessive 모델 하에서도 세 코호트 모두에서 확인되었다 (p 값은 각각 0.025, 0.013, 그리고 0.001). 복제 코호트에서는 dominant 모델 하에서 통계적 파워가 떨어졌지만 (p=0.028, 0.224, 그리고 0.013), 이는 적은 샘플 수 때문으로 추측된다. 동아시아인에서 *CRIM1* rs3821169 보유자는 높은 빈도를 보였고 (46.8%), 현 분석에서는 치오퓨린 독성에 대한 해당 변이의 homozygote 효과에 초점을 두고 진행하였다.

rs3821169 와 치오퓨린 독성 간 연관의 일관성을 평가하기 위하여, 일곱 개의 치오퓨린 독성 cutoff (*i.e.*, Group 1 (G1) ≤70%, G2 ≤60%, G3 ≤45%, G4 ≤35%, G5 ≤25%, 그리고 G6 <15% DIPs) 에서 후보 변이의 연관성이 평가되었다. 이 때 사용된 두 가지 대조군은 다음과 같다: (1) 발견, 복제, 그리고 혼합 군 각각 89, 21, 110 명으로 구성된 DIP > 70%을 보이는 백혈병 환자군 (G_0), 그리고 (2) 1000 Genomes Project 의 건강한 동아시아인 504 명으로 구성된 외부 대조군[55]. Dominant 와 recessive 모델 하에서 Fisher's exact test 와

Cochran-Armitage trend test (CATT)가 수행되었다. 6 개의 비교 군 중 4 개 그룹에서 Fisher's exact test (recessive 모델)와 CATTs 모두에서 G_0 그리고 외부 대조군 모두에 대해 통계적으로 유의한 연관성을 유지하였다 (표 14). Fluidigm genotyping 방법을 사용하여 blood sample 이 존재하는 118 명에 대해 rs3821169 의 genotype 을 확인한 결과 97.4%의 일치율을 보였다.

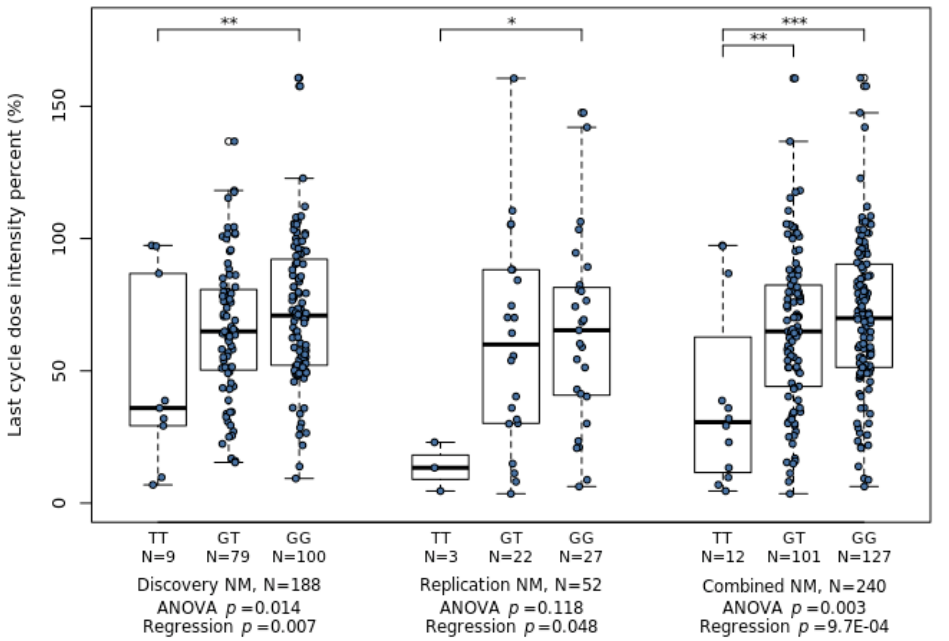


그림 15. *NUDT15* 과 *TPMT* 모두에 변이를 가지지 않는 소아 백혈병 환자에서 *CRIM1* rs3821169 변이와치오퓨린 약물 독성 간의 연관성. ANOVA 와 다중 선형 회귀 분석을 사용하여 *CRIM1* rs3821169 genotype 그룹 별로 발견 (p=0.014 그리고 0.007, N=188), 복제 (p=0.118 그리고 0.048, N=52), 그리고 혼합 (p=0.003 그리고 p=0.001, N=240) 코호트에서 유의한 6-MP 약물 강도 백분율 차이를 보이는 것을 확인함. *CRIM1*, gene encoding Cysteine-Rich Transmembrane BMP Regulator 1. *p<0.1, **p<0.05, ***p<0.01, post-hoc Tukey test.

표 14. 소아 급성 림프 모구성 백혈병 환자에서 허용되는 마지막 주기 6-MP 용량 강도 백분율의 다양한 역치에 걸친 *CRIM1* rs3821169 유전자형의 빈도 분포 평가.

		발굴 단계 (WES)								1000 Genomes EAS 와의 비교 (*N*=504)							
					dominant 모델		recessive 모델		CATT				dominant 모델		recessive 모델		CATT
종류	그룹	REF	HET	HOM	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*	그룹	REF	HET	HOM	*p*	OR (95% CI)	*p*	OR (95% CI)	*p*
혼합	G_0〉70	62	45	3						EAS	283	185	36					
	G_1≤70	65	56	9	0.364	1.29 (0.8-2.2)	0.234	2.64 (0.6-15.6)	0.168	G_1≤70	65	56	9	0.236	1.28 (0.9-1.9)	1.000	0.97 (0.4-2.1)	0.335
	G_2≤60	49	42	9	0.333	1.34 (0.8-2.4)	0.073	3.51 (0.8-20.7)	0.102	G_2≤60	49	42	9	0.226	1.33 (0.8-2.1)	0.532	1.29 (0.5-2.8)	0.192
	G_3≤45	19	26	9	**0.013**	2.37 (1.2-5.0)	**0.002**	7.04 (1.7-42.3)	**7.15E-04**	G_3≤45	19	26	9	**0.004**	2.36 (1.3-4.5)	**0.030**	2.59 (1-5.9)	**8.79E-04**
	G_4≤35	14	20	7	**0.018**	2.48 (1.1-5.7)	**0.004**	7.22 (1.5-45.6)	**0.001**	G_4≤35	14	20	7	**0.009**	2.47 (1.2-5.2)	**0.034**	2.67 (0.9-6.7)	**0.002**
	G_5≤25	8	8	5	0.154	2.09 (0.7-6.3)	**0.003**	10.8 (1.9-76.4)	**0.007**	G_5≤25	8	8	5	0.119	2.08 (0.8-5.9)	**0.018**	4.04 (1.1-12.4)	**0.014**
	G_6≤15	4	4	4	0.142	2.56 (0.6-12.3)	**0.002**	16.88 (2.4-136)	**0.003**	G_6≤15	4	4	4	0.145	2.56 (0.7-11.8)	**0.010**	6.45 (1.4-25.5)	**0.008**
발굴 단계	G_0〉70	51	35	3						EAS	283	185	36					
	G_1≤70	49	44	6	0.308	1.37 (0.7-2.5)	0.503	1.84 (0.4-11.7)	0.221	G_1≤70	49	44	6	0.227	1.31 (0.8-2.1)	0.831	0.84 (0.3-2.1)	0.416
	G_2≤60	37	31	6	0.430	1.34 (0.7-2.6)	0.302	2.52 (0.5-16.1)	0.202	G_2≤60	37	31	6	0.381	1.28 (0.8-2.2)	0.810	1.15 (0.4-2.9)	0.363
	G_3≤45	10	17	6	**0.014**	3.06 (1.2-8.1)	**0.012**	6.25 (1.2-41.3)	**0.001**	G_3≤45	10	17	6	**0.006**	2.94 (1.3-7.1)	**0.036**	2.88 (0.9-7.8)	**0.001**

	$G_4 \leq 35$	8	13	4	**0.040**	2.83 (1-8.4)	**0.040**	5.35 (0.8-39.4)	**0.006**	$G_4 \leq 35$	8	13	4	**0.023**	2.72 (1.1-7.4)	0.111	2.47 (0.6-7.9)	**0.011**
	$G_5 \leq 25$	3	4	2	0.292	2.66 (0.5-17.5)	0.065	7.85 (0.6-81.6)	**0.039**	$G_5 \leq 25$	3	4	2	0.193	2.56 (0.5-16)	0.138	3.7 (0.4-20.4)	0.074
	$G_6 \leq 15$	2	0	2	1	1.34 (0.1-19.2)	**0.013**	25.41 (1.4-472.8)	**0.049**	$G_6 \leq 15$	2	0	2	1	1.28 (0.1-17.8)	**0.030**	12.84 (0.9-181.8)	0.072
복제 단계	$G_0 > 70$	11	10	0						EAS	283	185	36					
	$G_1 \leq 70$	16	12	3	1	1.03 (0.3-3.6)	0.264	Inf (0.3-Inf)	0.604	$G_1 \leq 70$	16	12	3	0.710	1.2 (0.5-2.7)	0.486	1.39 (0.3-4.9)	0.543
	$G_2 \leq 60$	12	11	3	0.772	1.28 (0.3-4.7)	0.242	Inf (0.3-Inf)	0.369	$G_2 \leq 60$	12	11	3	0.321	1.49 (0.6-3.6)	0.428	1.69 (0.3-6)	0.256
	$G_3 \leq 45$	9	9	3	0.758	1.45 (0.4-5.9)	0.232	Inf (0.4-Inf)	0.246	$G_3 \leq 45$	9	9	3	0.266	1.71 (0.6-4.7)	0.199	2.16 (0.4-7.9)	0.146
	$G_4 \leq 35$	6	7	3	0.508	1.80 (0.4-8.5)	0.072	Inf (0.6-Inf)	0.128	$G_4 \leq 35$	6	7	3	0.200	2.13 (0.7-7.2)	0.111	2.99 (0.5-11.6)	0.059
	$G_5 \leq 25$	5	4	3	0.721	1.52 (0.3-8.3)	**0.040**	Inf (0.8-Inf)	0.141	$G_5 \leq 25$	5	4	3	0.384	1.79 (0.5-7.3)	0.055	4.31 (0.7-18.3)	0.080
	$G_6 \leq 15$	2	4	2	0.238	3.17 (0.4-39.2)	0.069	Inf (0.5-Inf)	**0.049**	$G_6 \leq 15$	2	4	2	0.147	3.83 (0.7-39.2)	0.113	4.31 (0.4-25.3)	**0.029**

REF: Reference; HET: Heterozygous; HOM: Homozygous; OR: Odds ratio; CATT: cochran armitage trend test; EAS: East Asian

4.3.4 치오퓨린 독성에 대한 *NUDT15*, *TPMT*, 그리고 *CRIM1* 의 복합 유전자 효과

기존에 잘 정립되어온 *NUDT15* 와 *TPMT* 대비 새로운 *CRIM1* 의 추가 효과를 평가하기 위하여, homozygous *CRIM1* rs3821169 효과를 반영하기 이전과 이후의 GVB 기반 ROC 분석을 진행하였다 (그림 16). 그림 16 은 전통적인 두개 유전자 기반 예측 모델 ($GVB^{NUDT15,TPMT}$)과 새로운 유전자를 포함한 세개 유전자 기반 예측 모델 ($GVB^{NUDT15,TPMT,CRIM1}$)의 진단적 정확도를 나타낸 AUC 이다. $GVB^{NUDT15,TPMT,CRIM1}$ 은 복제 코호트 의 DIP 〈 100% 기준을 제외하고는 (AUC〈100%=0.642 vs. 0.676) 모든 threshold cutoff 기준에서 전통적인 $GVB^{NUDT15,TPMT}$ 보다 더 좋은 예측력을 보였다 (*e.g.*, 각각 AUC〈15%=0.810 vs. 0.706, 0.697 vs. 0.600, 그리고 0.754 vs. 0.658; 각각 AUC〈25%=0.739 vs. 0.684, 0.728 vs. 0.633, 그리고 0.737 vs. 0.667).

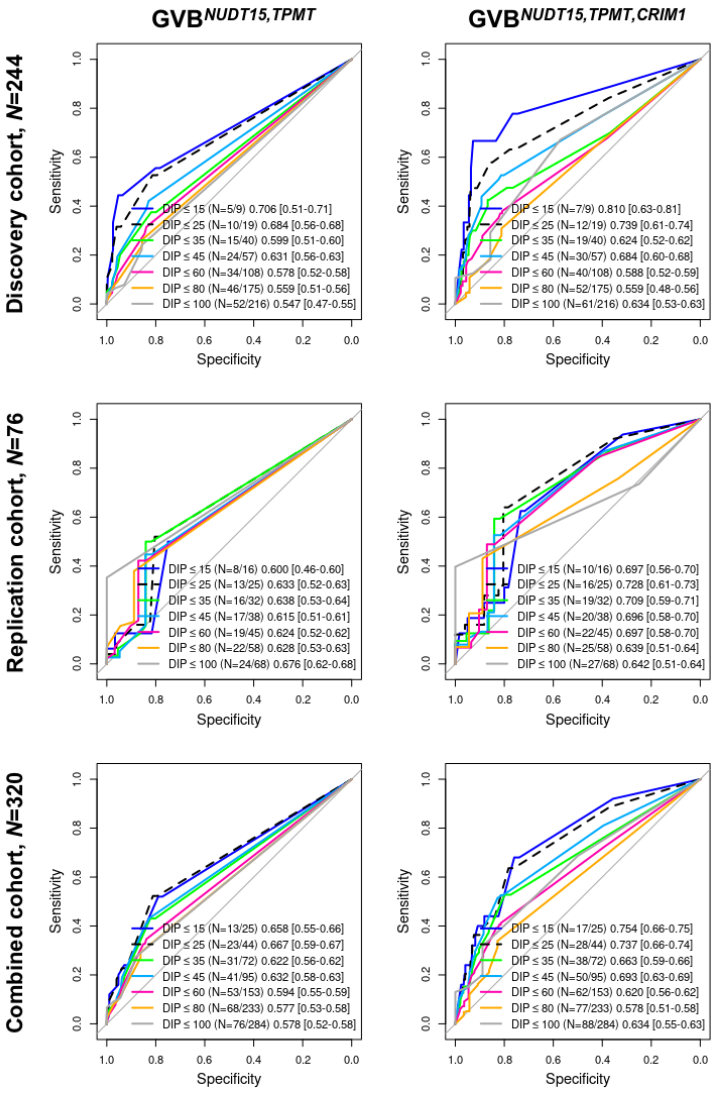


그림 16. 320 명의 소아 백혈병 환자에서 잘 확립 된 *NUDT15* 및 *TPMT* 에 *CRIM1* 을 도입함으로써 개선되는 약물 독성 예측 정확도. 일곱개의 DIP 역치 (≤15%, ≤25%, ≤35%, ≤45%, ≤60%, ≤80%, and ≤100%)를 기준으로 계산했을 때, 세 개 유전자 모델 (NUDT15, TPMT 및 CRIM1) (오른쪽 패널)의 마지막 주기 6-MP DIP 에 대한 예측 정확도 (AUC 에서 측정)는 기존의 두 가지 유전자 모델 (NUDT15 및 TPMT) (왼쪽 패널)보다 성능이 뛰어남. 95% 신뢰 구간이 괄호 안에 표기됨.

더 중요하게는, 6-MP 민감도를 예측하기 위한 용량-약물 반응성의 관계를 관찰할 수 있었는데, 낮은 DIP 일 수록 $GVB^{NUDT15,TPMT}$ 와 $GVB^{NUDT15,TPMT,CRIM1}$ 모두 더 높은 AUC 값을 가지는 것을 확인하였다. 예를 들어, 발견 단계의 $GVB^{NUDT15,TPMT,CRIM1}$ 의 경우, $AUC^{<15\%=}$0.810 는 $AUC^{<25\%}$=0.739 와 $AUC^{<35\%}$=0.624 보다 높았다. 추가로 *CRIM1* 의 단독 효과를 평가하기 위하여 GVB^{CRIM1}을 계산하였다. 동아시아인에서 높은 빈도를 보이는 rs3821169 의 특성을 반영하여, GVB^{CRIM1}는 heterozygous rs3821169 의 효과를 무시하고 계산하였다.

4.3.5 치오퓨린 독성에 대한 단일 유전자 효과

그림 17 은 전체 코호트 (320 명)에서 다른 두 유전자의 효과를 보정한 치오퓨린 독성에 대한 *CRIM1*, *NUDT15*, 그리고 *TPMT* 의 단일 유전자 예측 정확도를 나타낸다. GVB^{CRIM1} 의 AUC 는 *NUDT15* 과 TPMT 모두 WT 인 240 명의 환자에서 측정되었고, GVB^{NUDT15} 의 AUC 는 *TPMT* WT 이면서 *CRIM1* rs3821169 homozygote 변이를 가지고 있지 않은 294 명에서, 그리고 GVB^{TPMT} 의 AUC 는 *NUDT15* WT 이면서 *CRIM1* rs3821169 homozygote 변이를 가지고 있지 않은 236 명에서 측정되었다. 예측 정확도는 발견, 복제, 그리고 혼합 코호트 모두에서 평가되었다.

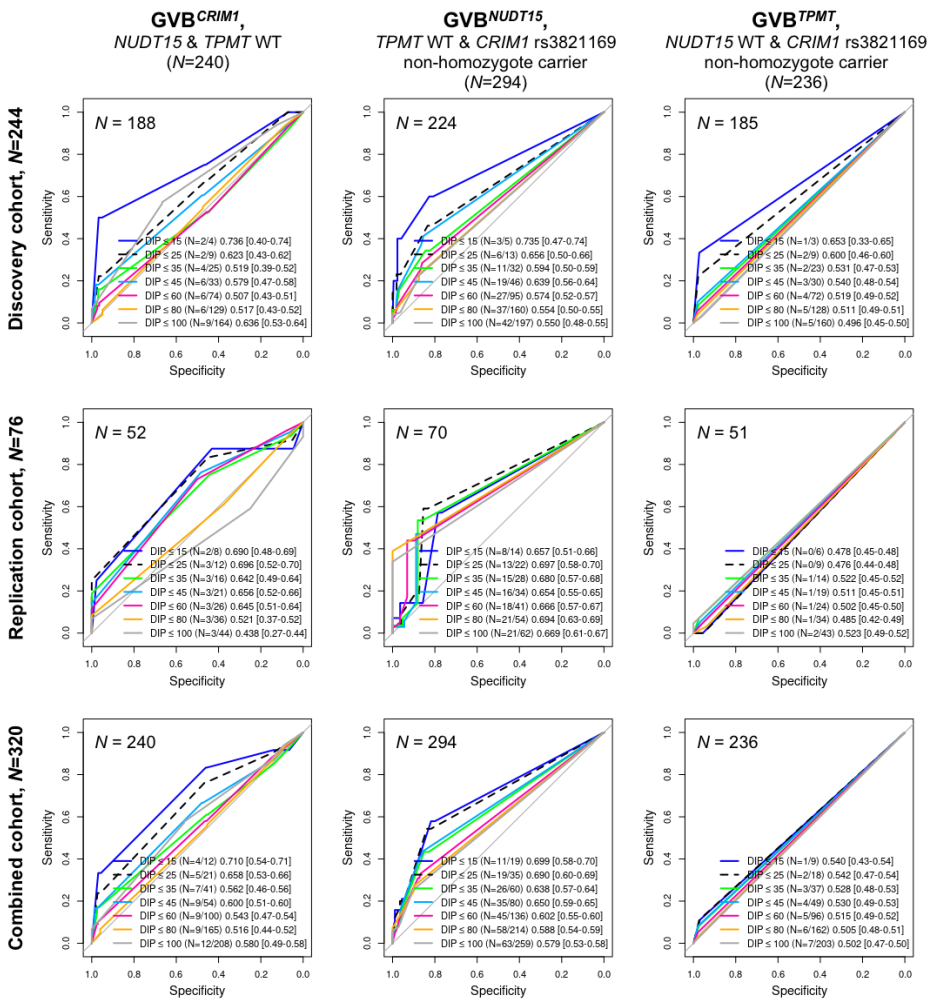


그림 17. 다른 두 유전자의 효과를 보정한 후 약물 독성 예측에의 *CRIM1*, *NUDT15* 및 *TPMT*의 단일 유전자 기여도 평가. 7 가지 컷오프(≤15 %, ≤25 %, ≤35 %, ≤45 %, ≤60 %, ≤80 % 및 ≤100)에서 다른 두 유전자의 효과를 제어한 뒤 마지막 사이클 6-MP DIP 를 예측하기 위한 GVB^{CRIM1}, GVB^{NUDT15} 및 GVB^{TPMT} 의 예측 정확도를 측정함. 95% 신뢰 구간은 괄호 안에 표기되어 있음.

전체적으로, *NUDT15*은 DIP < 25%의 cutoff에서 6-MP DIP에 대한 최고의 예측 정확도를 보였다 [발견 AUC=0.656, N=224; 복제 AUC=0.697, N=70; 혼합 AUC=0.690, N=294]. Recessive *CRIM1* model 은 동아시아인의 6-MP 민감도에 대한 가장 강력한 예측 지표인 *NUDT15*와 견줄만한 예측 정확도를 보였다 [발견 AUC=0.623, N=188; 복제 AUC=0.696, N=52; 혼합 AUC=0.658, N=240]. *TPMT* 는 현 분석에서는 예측력이 좋지 않았는데, 아마도 유럽인 대비 동양인에서 발견되는 빈도가 매우 낮을 것이기 때문으로 판단된다.

더 중요하게는, *NUDT15* 과 *CRIM1* 각각이 치오퓨린 독성 예측에 대한 용량-약물 반응성의 관계를 보였다는 것이다. 즉, *NUDT15* 과 *CRIM1* 모두 낮은 DIP threhold 기준일수록 높은 AUC 를 보였다. 종합적으로, 새로운 *CRIM1* 유전자는 homozygote 형태일 때 특히 높은 빈도로 발견되는 동아시아인에서 알려진 *NUDT15* 과 *TPMT* 유전자에 대해 독립적이고 추가적인 약물유전적 효과를 보이는 것으로 보인다.

4.3.6 *NUDT15*, *TPMT*, 그리고 *CRIM1*의 예측 정확도에 대한 평가

표 15 는 6-MP DIP 에 대한 *CRIM1* rs3821169 homozygote 의 예측 정확도를 나타낸다 [발견 (0.926), 복제 (0.827), 그리고 혼합 (0.904) 코호트]. *CRIM1* rs3821169 변이는 그 자체로는 상대적으로 낮은 민감도 (0.222~0.250)와 양성 예측도 (0.222~1.000), 그리고 상대적으로 높은 특이도 (0.961~1.000)와 음성 예측도 (0.816~0.961)를 갖는다.

현재의 6-MP 에 대한 CPIC 가이드라인은 *TPMT* 와 *NUDT15* 에 대한 스타 대립 유전자 기반의 diplotype 을 적용하여 평가한다. 스타 대립 유전자는 genotype 의 세트로부터 추정되는데, CPIC 가이드 라인은 일반적으로 범주형 대립 유전자 클래스에 대해 다중 유전자 상호 작용을 결합하는 방법에 대한 특정 지침을 제공하지 않는다. 또한, *CRIM1* 의 경우 아직 스타 대립 유전자가 부여되지 않았기 때문에, 여러 유전자의 약리학적 테스트를 적용하여 그 임상적 유용성을 평가하는 것은 매우 중요한 문제로 남아있다. 여러 유전자의 효과를 종합하는 GVB 점수의 유용성을 평가하기 위하여, 전통적인 스타 대립 유전자 기반의 *NUDT15* 과 *TPMT* 의 진단적 정확도를 체계적으로 비교 평가 하였다 (표 16). 이 때, GVB 점수의 적절한 cutoff 값은 Youden's index 를 최대화시키는 GVB 값으로 설정되었다 (그림 18).

표 15. 약물 독성 예측에 있어서 *CRIM1* rs3821169 변이의 예측 정확도 평가.

단계	rs3821169 homozygote 보유자	6-MP DIP (%) ≤25%	>25%	전체	민감도	특이도	양성 예측도	음성 예측도	정확도
발견	(+)	2	7	9	0.222	0.961	0.222	0.961	0.926
	(−)	7	172	179					
	전체	9	179	188					
복제	(+)	3	0	3	0.250	1.000	1.000	0.816	0.827
	(−)	9	40	49					
	전체	12	40	52					
혼합	(+)	5	7	12	0.238	0.968	0.417	0.930	0.904
	(−)	16	212	228					
	전체	21	219	240					

표 16. 소아 백혈병 환자에서 약물 독성을 예측하기 위한 스타 대립 유전자 기반 분자 표현형 대 유전자 수준 변이 부담 점수의 정확도 비교 평가. *NUDT15* 및 *TPMT*에 대한 스타 대립 유전자 기반 CPIC 가이드 라인의 마지막 사이클 6-MP DIP에 대한 예측 정확도를 발견, 복제, 혼합군에서의 정량적 $GVB^{NUDT15,TPMT}$ 과 $GVB^{NUDT15,TPMT,CRIM1}$ 방법과 비교. GVB 임계값은 Youden 의 지수를 최대화하는 값으로 결정됨.

단계	방법론	분자 표현형	6-MP DIP			민감도	특이도	양성 예측도	음성 예측도	정확도
			≤25%	>25%	전체					
발굴	CPIC *NUDT15* 과 *TPMT* metabolizer	PM+IM	10	46	56	0.526	0.796	0.179	0.952	0.775
		NM	9	179	188					
	$GVB^{NUDT15,TPMT}$	≤0.3	10	42	52	0.526	0.813	0.192	0.953	0.791
		>0.3	9	183	192					
	$GVB^{NUDT15,TPMT,CRIM1}$	≤0.3	11	32	43	0.579	0.858	0.256	0.960	0.836
		>0.3	8	193	201					
		전체	19	225	244					
복제	CPIC *NUDT15* 과 *TPMT* metabolizer	PM+IM	13	11	24	0.520	0.784	0.542	0.769	0.697
		NM	12	40	52					
	$GVB^{NUDT15,TPMT}$	≤0.3	13	10	23	0.520	0.804	0.565	0.774	0.711
		>0.3	12	41	53					

	GVB^{NUDT15,TPMT,CRIM1}	≤0.45	16	10	26	0.640	0.804	0.615	0.820	0.750
		>0.45	9	41	50					
		전체	25	51	76					
혼합	CPIC *NUDT15* 과 *TPMT* metabolizer	PM+IM	23	57	80	0.523	0.794	0.288	0.913	0.756
		NM	21	219	240					
	GVB^{NUDT15,TPMT}	≤0.3	23	52	75	0.523	0.811	0.307	0.914	0.772
		>0.3	21	224	245					
	GVB^{NUDT15,TPMT,CRIM1}	≤0.45	28	60	88	0.636	0.783	0.318	0.931	0.763
		>0.45	16	216	232					
		전체	44	276	320					

IM, intermediate metabolizer; PM, poor metabolizer

그림 18. $GVB^{NUDT15,TPMT}$ and $GVB^{NUDT15,TPMT,CRIM1}$의 적정 임계값을 찾기 위한 Youden's index 계산 결과.

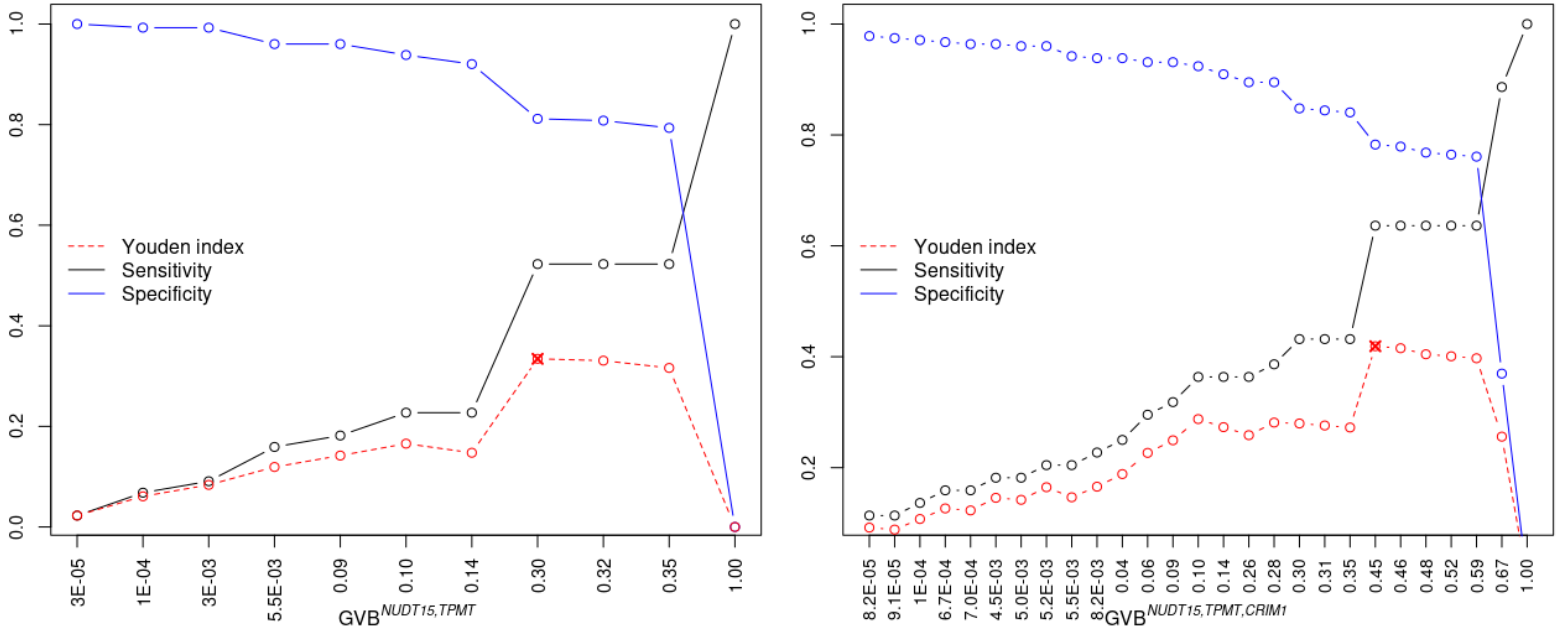


표 16 을 통하여 GVB^{NUDT15,TPMT}가 민감도, 특이도, 양성 예측도, 음성 예측도를 증가시키면서 기존의 스타 대립 유전자 기반의 분류보다 약간 더 나은 예측 정확도를 보이는 것을 확인할 수 있다 [발견 (0.791 vs. 0.775), 복제 (0.711 vs. 0.697), 그리고 혼합 (0.772 vs. 0.756)]. *CRIM1* 의 경우 아직 지정된 스타 대립 유전자가 존재하지 않기 때문에, GVB 기반 세 유전자 예측 모델을 생성하였다: GVB^{NUDT15,TPMT,CRIM1} 는 전통적인 스타 대립 유전자 기반의 *NUDT15* 과 *TPMT* diplotyping 방법론에 비해 높은 예측 정확도를 보였다 [발견 (0.836 vs. 0.775), 복제 (0.750 vs. 0.697), 그리고 혼합 (0.763 vs. 0.756)]. GVB^{NUDT15,TPMT,CRIM1}는 GVB^{NUDT15,TPMT} 와 비교했을 때도 높은 예측 정확도를 보였는데, 발견, 복제, 그리고 혼합 군 각각에서 높은 민감도 (0.579 vs. 0.526, 0.640 vs. 0.520, 그리고 0.636 vs. 0.523, respectively), 양성 예측도 (0.256 vs. 0.192, 0.615 vs. 0.563, 그리고 0.318 vs. 0.307), 그리고 음성 예측도 (0.960 vs. 0.953, 0.820 vs. 0.774, 0.931 vs. 0.914)를 보였다. 특이도 (0.858 vs. 0.813, 0.804 vs. 0.804, 0.783 vs. 0.811) 와 정확도 (0.836 vs. 0.791, 0.750 vs. 0.711, 그리고 0.763 vs. 0.772)는 발견과 복제 군에서는 증가하였으나, 혼합 군에서는 약간 감소하였다.

4.4 고찰

*CRIM1*은 골 형성 단백질과 상호작용하는 것으로 알려진 발달상 중요한 단백질 (BMPs, bone morphogenetic proteins)과 유사한 세포 표면 막통과 (cell-surface transmembrane) 단백질이다. 약물 저항성에 대한 *CRIM1* 의 역할은 이전 여러 연구에서 밝혀진 적 있는데 [88, 89], *CRIM1* 의 mRNA expression level 이 높은 경우 leukemic cell 에 대한 저항성을 보일 수 있다는 보고가 있다. 이것은 BMP level 에 영향을 미치고, 따라서 *CRIM1* 이 조혈세포의 성장과 분화를 조절한다는 것을 시사한다. GDSC (Genomics of Drug Sensitivity in Cancer) 자료 [90]를 활용하여 rs3821169 heterozygous 보유군이 WT 군 대비 낮은 mRNA expression level 을 보인다는 것을 확인하였다 (그림 19, *t*-test p=0.095). 서양인에서 해당 변이의 상대적 빈도가 매우 낮기때문에 homozygote 변이 보류자가 발견되지 않아 상응하는 단백질의 기능 상실 (loss of function) 가능성을 예측할 수는 없었지만, 해당 변이가 약물-민감성 반응에 어느 정도 영향을 미칠 가능성이 있음을 나타낸다. *CRIM1* 이 치오퓨린 독성에 어떻게 영향을 미칠 수 있을지에 대한 실험적 입증이 추가로 필요하다.

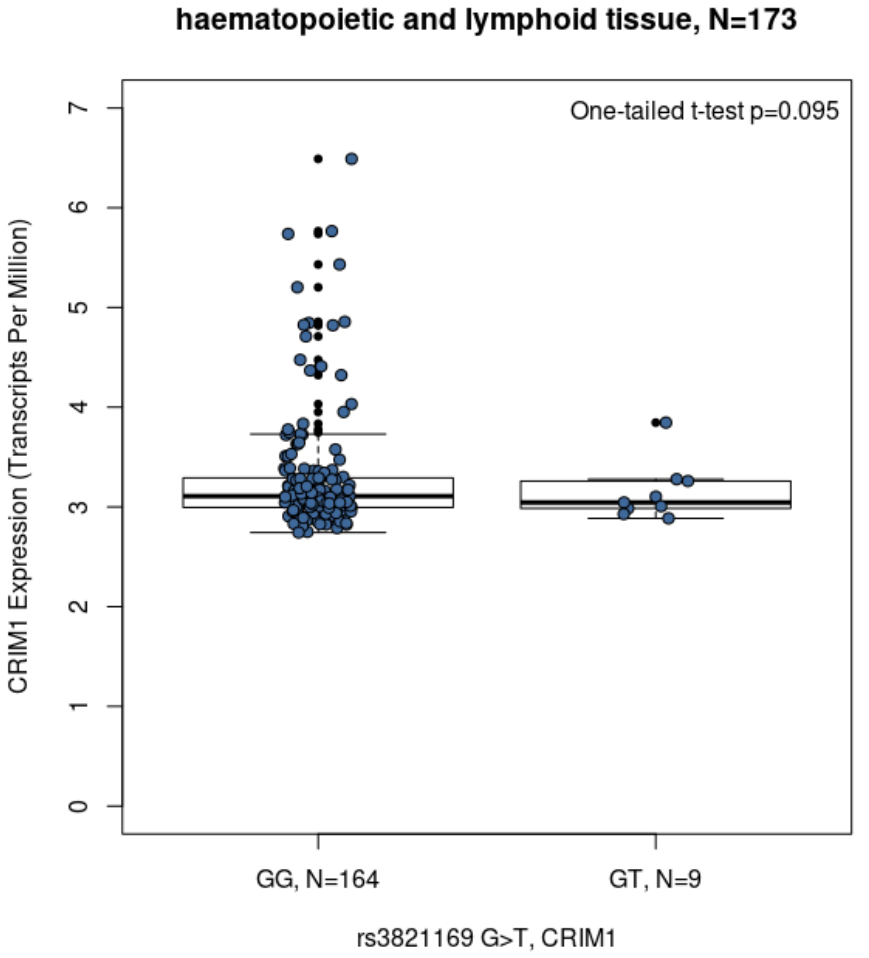


그림 19. 조혈 및 림프 조직에서 rs3821169 변이 보유군과 비보유군 간 *CRIM1* mRNA expression levels 의 비교.

현 분석에서는 *CRIM1*을 소아 백혈병 환자에서 치오퓨린 약물 독성을 예측할 새로운 인자로 제안한다. *CRIM1* rs3821169는 잠재적으로 해롭고 (SIFT score=0, CADD score=25.3), 동아시아 집단에서 매우 빈번하게 발견되기 때문에 (MAF=25%) 분석 예측력을 증가시켰다. 높은 대립 유전자 빈도로부터 예상할 수 있듯, homozygous 모델은 6-MP 민감도에 대한 예측 정확도를 증가시켰으나, heterozygous 모델은 phenotypic effect에 중간 정도의 (moderate) 효과를 보였다.

CRIM1 rs3821169 의 빈도는 동아시아인에서 (T=0.255) 다른 인종보다 (global=0.066, Africans=0.001, Europeans=0.009, South Asians=0.05, 그리고 Americans=0.02; 1000 Genomes Project, Phase 3) 현저히 높았다. Homozygous 보유자는 동아시아인에서만 발견되었는데 (*T*=0.071), 이러한 높은 인종별 분포 차이가 해당 변이가 아직 치오퓨린 독성의 biomarker로 발견되지 않은 이유를 일부 설명할 수 있을 것으로 보인다. 현재의 연구는 유럽인종에 상당히 치우쳐져 있기 때문에 [91], 해당 변이에 대한 통계적 파워를 얻기에는 충분하지 않았을 가능성이 있다. 현 분석에는 6-MP 유지요법을 진행한 상대적으로 많은 동아시아인 (320 명의 한국인)을 포함하고 있기 때문에 이미 잘 알려진 강력한 인자 (*NUDT15*과 *TPMT*)의 영향을 제외하고 설정한 both WT 군을 대상으로 새로운 biomarker를 탐색하는 것이 가능했다.

약물유전체 변이의 높은 인종 간 빈도 차이는 매우 주목할 만하다. 최근 발견된 치오퓨린 독성 연관 매우 강력한 유전 인자인 *NUDT15* rs116855232 변이 또한 동아시아인에서 (*T*=0.095) 다른 인종 대비 (global=0.040, Africans=0.001, Europeans=0.002, South Asians=0.07, 그리고 Americans=0.04; 1000 Genomes Project, Phase 3) 매우 높은 빈도로 발견된다 [92]. 정의상 약물 유전자의 경우 질병 연관 유전자와 달리 '약'이라는 외부 자극 없이는 특정 표현형을 보이지 않고, 이렇게 약물 유전자에서 명백한 표현형이 나타나지 않는 것은 상이한 환경에서 다양한 진화적 선택 압력 (selective pressure) 하에 높은 인종 간 다양성을 허용하게 했을 수 있기 때문으로 추측된다.

결론적으로, *CRIM1* 은 6-MP 유발 약물 독성과 연관된 유전자이다. 현 스터디에서 제안한 증거들은 genome-wide 한 유의성을 확보하기에 불충분한 샘플 수, 그리고 인종적 다양성 등의 한계를 갖는다. 추후 분석에서는 6-MP 대사에서의 *CRIM1* 의 역할을 규명할 필요가 있다.

5 고찰

본 연구에서는 임상에서 평가적 방법론으로서 유전자 수준 변이 부담 점수를 얼마나 적절히 활용할 수 있을 지에 대하여 평가하였다. 유전자 변이 부담 점수는 특정 질병이나 특성에 대한 유전적 위험도를 개인에 대한 점수로 환산하여 제공하고 이를 통해 고위험군의 환자를 탐지한다는 점에서 기존의 유전자 기반 통계 테스트나 인구집단 기반의 점수 체계들과는 매우 상이한 목적을 가지고 있다. 특히, 유전자 수준 변이 부담 점수는 개인에게서 발견된 변이들의 위해도를 진화적 압력에 기반하여 예측하고 이를 효과적으로 통합함으로써 매우 가변적인 영역(variable region)에 위치하는 위험 변인들을 예측하는 데 좋은 성능을 보이는 특징을 가지고 있다 (*i.e.*, 약물 유전자). 이는 매우 보존된 영역에서 발생한 변이를 탐지하는 데 초점을 두고 개발되었던 기존의 인구 집단 기반의 유전자 점수들 (*i.e.*, 희귀 질환 유전자)과 상호 보완적인 역할을 할 수 있다는 점에서 그 활용에 이점이 존재한다.

유전자 수준 변이 부담 점수는 평가적 방법론 뿐 아니라 탐색적 방법론으로서도 활용되고 있다. 최근 bisphosphonate 관련 악골 괴사증 환자에서 거짓 음성 결과 (false negatives)를 최소화하기 위한 목적으로 유전자 수준 변이 부담 점수와 통계 테스트 (SKAT-O 와 burden test)를

동시에 적용한 경우가 있었는데 [18], 아직 기존의 통계 테스트 대비 탐색적 접근으로서의 유전자 수준 변이 부담 점수의 성능에 대한 벤치마크 테스트가 체계적으로 진행된 적이 없기 때문에 해당 점수를 탐색적 접근 방법으로 활용하기 위해서는 정교한 추가 검증이 필요할 것이다.

기존의 인구 집단 기반 점수들과 비교했을 때, 유전자 수준 변이 부담 점수는 약물 유전자와 비슷한 유전적 조성을 가지는 유전자들을 예측하는 데 뛰어난 성능을 보이는 것을 확인하였다. 일곱가지 유전적 특성 (genetic feature)을 기반으로 희귀 질환, 복합 질환, 그리고 약물 유전자에 대한 특성화 (characterization)를 진행한 결과를 유전자 중심의 개념으로 시각화 해보았다 (그림 20). 희귀질환 유전자는 매우 보존된 영역에서 (x 축) 높은 penetrance 를 보이며(y 축), 강력한 효과를 가지는 (원의 색) 소수의 변이 (원의 크기)에 의해 설명되는 특성을 가지는 반면, 복합질환은 가변적인 영역에서 상대적으로 낮은 설명력을 가지는 낮은 효과의 여러 변이에 의해 설명되는 특성을, 약물 유전자는 음식에 대한 방어체계로 지역적/인종적 특성을 갖지만 비교적 중등도 이상의 선택적 압력을 가지는 등, 각 표현형 별로 매우 상이한 특징을 가지고 있다. 이렇게 표현형 별 특이적인 유전자의 특성을 파악하는 것은 추후 관심 있는 표현형과 연관된 유전자 우선순위를 결정하는 데 도움이 될 것이다.

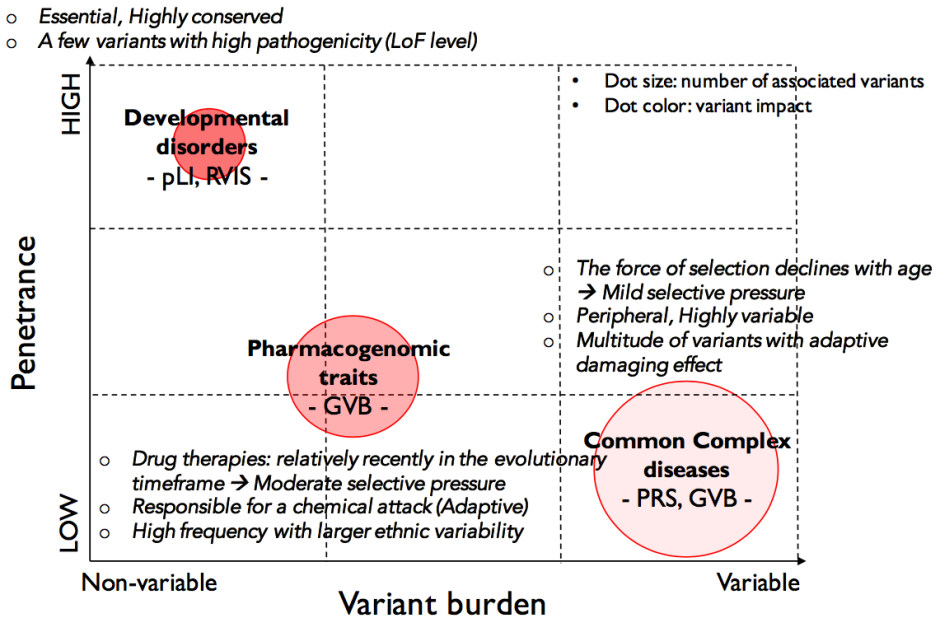


그림 20. 약물, 희귀질환, 복합질환 유전자에 대한 유전자 중심의 특성 분포

요약하면, 유전자 수준 변이 부담 점수는 특히 약물 유발 부작용과 관련된 유전자를 활용하여 환자의 위험도를 계층화하고 평가하는 데 도움이 되는 방법론이다. 방법론의 가치에 대한 명확한 증거를 제공하기 위해서 보다 확장된 분야에서 유전자 수준 변이 부담 점수의 유용성과 역할에 대한 지속적인 연구와 구현이 필요할 것이다.

참고문헌

1. International Hapmap C, Altshuler DM, Gibbs RA *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* 467(7311), 52-58 (2010).
2. Genomes Project C, Abecasis GR, Auton A *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 491(7422), 56-65 (2012).
3. Elsharawy A, Warner J, Olson J *et al.* Accurate variant detection across non-amplified and whole genome amplified DNA using targeted next generation sequencing. *BMC Genomics* 13 500 (2012).
4. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19(8), 491-504 (2018).
5. Gorlov IP, Gorlova OY, Frazier ML, Spitz MR, Amos CI. Evolutionary evidence of the effect of rare variants on disease etiology. *Clin Genet* 79(3), 199-206 (2011).
6. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19(3), 212-219 (2009).
7. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol* 18(1), 77 (2017).
8. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. *American journal of human genetics* 95(1), 5-23 (2014).
9. Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* 9(8), e1003709 (2013).
10. Itan Y, Shang L, Boisson B *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci U S A* 112(44), 13615-13620 (2015).
11. Segura-Lepe MP, Keun HC, Ebbels TMD. Predictive modelling using pathway scores: robustness and significance of pathway collections. *BMC bioinformatics* 20(1), (2019).

12. Gussow AB, Petrovski S, Wang Q, Allen AS, Goldstein DB. The intolerance to functional genetic variation of protein domains predicts the localization of pathogenic mutations within genes. *Genome Biol* 17 9 (2016).
13. Lek M, Karczewski KJ, Minikel EV *et al*. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616), 285-291 (2016).
14. Lee KH, Baik SY, Lee SY, Park CH, Park PJ, Kim JH. Genome Sequence Variability Predicts Drug Precautions and Withdrawals from the Market. *PloS one* 11(9), e0162135 (2016).
15. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* 19(9), 581-590 (2018).
16. Park Y, Kim H, Choi JY *et al*. Star Allele-Based Haplotyping versus Gene-Wise Variant Burden Scoring for Predicting 6-Mercaptopurine Intolerance in Pediatric Acute Lymphoblastic Leukemia Patients. *Frontiers in pharmacology* 10 654 (2019).
17. Park J, Lee SY, Baik SY *et al*. Gene-Wise Burden of Coding Variants Correlates to Noncoding Pharmacogenetic Risk Variants. *Int J Mol Sci* 21(9), (2020).
18. Lee KH, Kim SH, Kim CH *et al*. Identifying genetic variants underlying medication-induced osteonecrosis of the jaw in cancer and osteoporosis: a case control study. *J Transl Med* 17(1), 381 (2019).
19. Seo H, Kwon EJ, You YA *et al*. Deleterious genetic variants in ciliopathy genes increase risk of ritodrine-induced cardiac and pulmonary side effects. *BMC medical genomics* 11(1), 4 (2018).
20. Orr HA. Fitness and its role in evolutionary genetics. *Nat Rev Genet* 10(8), 531-539 (2009).
21. Vogenberg FR, Isaacson Barash C, Pursel M. Personalized medicine: part 1: evolution and development into theranostics. *P T* 35(10), 560-576 (2010).
22. Dubinsky MC, Lamothe S, Yang HY *et al*. Pharmacogenomics and metabolite measurement for 6-mercaptopurine therapy in inflammatory bowel disease. *Gastroenterology* 118(4), 705-713 (2000).
23. Ansari A, Hassan C, Duley J *et al*. Thiopurine methyltransferase

activity and the use of azathioprine in inflammatory bowel disease. *Aliment Pharmacol Ther* 16(10), 1743-1750 (2002).
24. Cuffari C. A Physician's Guide to Azathioprine Metabolite Testing. *Gastroenterol Hepatol (N Y)* 2(1), 58-63 (2006).
25. Bradford K, Shih DQ. Optimizing 6-mercaptopurine and azathioprine therapy in the management of inflammatory bowel disease. *World J Gastroenterol* 17(37), 4166-4173 (2011).
26. Supandi S, Harahap Y, Harmita H, Andalusia R. Quantification of 6-Mercaptopurine and Its Metabolites in Patients with Acute Lympoblastic Leukemia Using Dried Blood Spots and UPLC-MS/MS. *Sci Pharm* 86(2), (2018).
27. Gonzalez-Lama Y, Gisbert JP. Monitoring thiopurine metabolites in inflammatory bowel disease. *Frontline Gastroenterol* 7(4), 301-307 (2016).
28. Lennard L. Implementation of TPMT testing. *Br J Clin Pharmacol* 77(4), 704-714 (2014).
29. Kakuta Y, Kinouchi Y, Shimosegawa T. Pharmacogenetics of thiopurines for inflammatory bowel disease in East Asia: prospects for clinical application of NUDT15 genotyping. *J Gastroenterol* 53(2), 172-180 (2018).
30. Yang JJ, Landier W, Yang WJ *et al*. Inherited NUDT15 Variant Is a Genetic Determinant of Mercaptopurine Intolerance in Children With Acute Lymphoblastic Leukemia. *Journal of Clinical Oncology* 33(11), 1235-+ (2015).
31. Zgheib NK, Akika R, Mahfouz R *et al*. NUDT15 and TPMT genetic polymorphisms are related to 6-mercaptopurine intolerance in children treated for acute lymphoblastic leukemia at the Children's Cancer Center of Lebanon. *Pediatr Blood Cancer* 64(1), 146-150 (2017).
32. Relling MV, Gardner EE, Sandborn WJ *et al*. Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. *Clin Pharmacol Ther* 93(4), 324-325 (2013).
33. Relling MV, Schwab M, Whirl-Carrillo M *et al*. Clinical Pharmacogenetics Implementation Consortium Guideline for Thiopurine Dosing Based on TPMT and NUDT15 Genotypes: 2018 Update. *Clin Pharmacol Ther* 105(5), 1095-1105 (2019).

34. Stephens M, Scheet P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *American journal of human genetics* 76(3), 449-462 (2005).
35. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *American journal of human genetics* 68(4), 978-989 (2001).
36. Whirl-Carrillo M, Mcdonagh EM, Hebert JM *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* 92(4), 414-417 (2012).
37. Moriyama T, Nishii R, Perez-Andreu V *et al.* NUDT15 polymorphisms alter thiopurine metabolism and hematopoietic toxicity. *Nature genetics* 48(4), 367-373 (2016).
38. Kim H, Kang HJ, Kim HJ *et al.* Pharmacogenetic analysis of pediatric patients with acute lymphoblastic leukemia: a possible association between survival rate and ITPA polymorphism. *PloS one* 7(9), e45558 (2012).
39. Robin X, Turck N, Hainard A *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC bioinformatics* 12 77 (2011).
40. Shah RR, Shah DR. Personalized medicine: is it a pharmacogenetic mirage? *Brit J Clin Pharmaco* 74(4), 698-721 (2012).
41. Moriyama T, Yang YL, Nishii R *et al.* Novel variants in NUDT15 and thiopurine intolerance in children with acute lymphoblastic leukemia from diverse ancestry. *Blood* 130(10), 1209-1212 (2017).
42. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* 46(3), 310-315 (2014).
43. Dong CL, Wei P, Jian XQ *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human Molecular Genetics* 24(8), 2125-2137 (2015).
44. Beyene J, Tritchler D, Asimit JL, Hamid JS. Gene- or Region-Based Analysis of Genome-Wide Association Studies. *Genetic Epidemiology* 33 S105-S110 (2009).
45. Grimm DG, Azencott CA, Aicheler F *et al.* The Evaluation of Tools

Used to Predict the Impact of Missense Variants Is Hindered by Two Types of Circularity. *Human Mutation* 36(5), 513–523 (2015).

46. Miosge LA, Field MA, Sontani Y *et al.* Comparison of predicted and actual consequences of missense mutations. *Proc Natl Acad Sci U S A* 112(37), E5189–5198 (2015).
47. Starita LM, Ahituv N, Dunham MJ *et al.* Variant Interpretation: Functional Assays to the Rescue. *American journal of human genetics* 101(3), 315–325 (2017).
48. Stessman HA, Bernier R, Eichler EE. A genotype–first approach to defining the subtypes of a complex disease. *Cell* 156(5), 872–877 (2014).
49. Blekhman R, Man O, Herrmann L *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr Biol* 18(12), 883–889 (2008).
50. Barreiro LB, Laval G, Quach H, Patin E, Quintana–Murci L. Natural selection has driven population differentiation in modern humans. *Nature genetics* 40(3), 340–345 (2008).
51. Hovelson DH, Xue Z, Zawistowski M *et al.* Characterization of ADME gene variation in 21 populations by exome sequencing. *Pharmacogenet Genomics* 27(3), 89–100 (2017).
52. Li J, Zhang L, Zhou H, Stoneking M, Tang K. Global patterns of genetic diversity and signals of natural selection for human ADME genes. *Hum Mol Genet* 20(3), 528–540 (2011).
53. Landrum MJ, Lee JM, Riley GR *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* 42(Database issue), D980–985 (2014).
54. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci Rep* 8(1), 1362 (2018).
55. Genomes Project C, Auton A, Brooks LD *et al.* A global reference for human genetic variation. *Nature* 526(7571), 68–74 (2015).
56. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high–throughput sequencing data. *Nucleic acids research* 38(16), e164 (2010).
57. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome research* 11(5), 863–874 (2001).
58. Ng PC, Henikoff S. Accounting for human polymorphisms

predicted to affect protein function. *Genome research* 12(3), 436–446 (2002).
59. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* 31(13), 3812–3814 (2003).
60. Adzhubei IA, Schmidt S, Peshkin L *et al.* A method and server for predicting damaging missense mutations. *Nature Methods* 7(4), 248–249 (2010).
61. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research* 20(1), 110–121 (2010).
62. Schwarz JM, Rodelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature Methods* 7(8), 575–576 (2010).
63. Cooper GM, Stone EA, Asimenos G *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome research* 15(7), 901–913 (2005).
64. Gong L, Owen RP, Gor W, Altman RB, Klein TE. PharmGKB: an integrated resource of pharmacogenomic data and knowledge. *Curr Prot Bioinformatics* Chapter 14 Unit14–17 (2008).
65. Wishart DS, Knox C, Guo AC *et al.* DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic acids research* 36(Database issue), D901–906 (2008).
66. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. *Nature genetics* 36(5), 431–432 (2004).
67. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, Mckusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic acids research* 30(1), 52–55 (2002).
68. Bartha I, Di Iulio J, Venter JC, Telenti A. Human gene essentiality. *Nat Rev Genet* 19(1), 51–62 (2018).
69. Hermjakob H, Montecchi-Palazzi L, Lewington C *et al.* IntAct: an open source molecular interaction database. *Nucleic acids research* 32(Database issue), D452–455 (2004).
70. Doyle MA, Gasser RB, Woodcroft BJ, Hall RS, Ralph SA. Drug target prediction and prioritization: using orthology to predict essentiality in parasite genomes. *BMC Genomics* 11 222 (2010).

71. Gu X. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet* 19(7), 354–356 (2003).
72. Conant GC, Wagner A. Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. *Proc Biol Sci* 271(1534), 89–96 (2004).
73. Jin W, Qin P, Lou H, Jin L, Xu S. A systematic characterization of genes underlying both complex and Mendelian diseases. *Hum Mol Genet* 21(7), 1611–1624 (2012).
74. Eyre-Walker YC, Eyre-Walker A. The role of mutation rate variation and genetic diversity in the architecture of human disease. *PloS one* 9(2), e90166 (2014).
75. LóPez-Bigas N, Ouzounis CA. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic acids research* 32(10), 3108–3114 (2004).
76. Mcdonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351(6328), 652 (1991).
77. Osada N, Mano S, Gojobori J. Quantifying dominance and deleterious effect on human disease genes. *Proc Natl Acad Sci U S A* 106(3), 841–846 (2009).
78. Lu HC, Chung SS, Fornili A, Fraternali F. Anatomy of protein disorder, flexibility and disease-related mutations. *Front Mol Biosci* 2 47 (2015).
79. Walter J, Charon J, Hu Y *et al.* Comparative analysis of mutational robustness of the intrinsically disordered viral protein VPg and of its interactor eIF4E. *PloS one* 14(2), e0211725 (2019).
80. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 17(2), 149–163 (1993).
81. Shakhnovich BE, Koonin EV. Origins and impact of constraints in evolution of gene families. *Genome research* 16(12), 1529–1536 (2006).
82. Chen WH, Zhao XM, Van Noort V, Bork P. Human monogenic disease genes have frequently functionally redundant paralogs. *PLoS Comput Biol* 9(5), e1003073 (2013).
83. Schaeffeler E, Jaeger SU, Klumpp V *et al.* Impact of NUDT15 genetics on severe thiopurine-related hematotoxicity in patients with European ancestry. *Genet Med* 21(9), 2145–2150 (2019).

84. Relling MV, Klein TE. CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clin Pharmacol Ther* 89(3), 464-467 (2011).
85. Yang JJ, Landier W, Yang W *et al.* Inherited NUDT15 variant is a genetic determinant of mercaptopurine intolerance in children with acute lymphoblastic leukemia. *J Clin Oncol* 33(11), 1235-1242 (2015).
86. Kim H, Seo H, Park Y *et al.* APEX1 Polymorphism and Mercaptopurine-Related Early Onset Neutropenia in Pediatric Acute Lymphoblastic Leukemia. *Cancer Res Treat* 50(3), 823-834 (2018).
87. Cingolani P, Platts A, Wang Le L *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* 6(2), 80-92 (2012).
88. Prenkert M, Uggla B, Tidefelt U, Strid H. CRIM1 is expressed at higher levels in drug-resistant than in drug-sensitive myeloid leukemia HL60 cells. *Anticancer Res* 30(10), 4157-4161 (2010).
89. Ziliak D, Gamazon ER, Lacroix B, Kyung Im H, Wen Y, Huang RS. Genetic variation that predicts platinum sensitivity reveals the role of miR-193b* in chemotherapeutic susceptibility. *Mol Cancer Ther* 11(9), 2054-2061 (2012).
90. Iorio F, Knijnenburg TA, Vis DJ *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* 166(3), 740-754 (2016).
91. Sirugo G, Williams SM, Tishkoff SA. The Missing Diversity in Human Genetic Studies. *Cell* 177(4), 1080 (2019).
92. Yang SK, Hong M, Baek J *et al.* A common missense variant in NUDT15 confers susceptibility to thiopurine-induced leukopenia. *Nature genetics* 46(9), 1017-1020 (2014).

Abstract

Genomic Characterization of Pharmacogenomic and Disease Genes using Gene-wise Variant Burden: Evidence of Utility in the Field of Computational Pharmacogenomics

Yoomi Park
The Department of Bioinformatics
The Graduate School
Seoul National University

The advent of next-generation sequencing technologies has empowered researchers with the ability to catalogue and predict the contribution of many different types of clinically relevant genetic variants. The traditional single variant-based analysis is limited since the rarity limits the statistical power of associating rare variants with phenotypes, requiring a large sample size. To alleviate this problem, gene-based (or region-based) approaches that aggregate the impact of multiple variants in a gene (or a region) have been proposed. The recently published Gene-wise Variant Burden (GVB) score, a score that integrates the overall deleterious impacts of multiple variants on a gene in an individual-specific manner, has been previously utilized in the field of pharmacogenetics, but the utility of the

score has not been systematically evaluated. In this study, a comprehensive evaluation of the utility of GVB was performed in translating genotype information into phenotype across PGx, complex-disease, and Mendelian-disease genes.

GVB scores were computed and assigned for protein-coding genes for each of the 2504 individual genomes from the 1000 Genomes Project (1KGP) and 320 pediatric acute lymphoblastic leukemia (ALL) patients. To assess the utility of GVB scoring method in quantifying the potential contributing effect of variants on enzymatic activity, we performed a comparison study of the conventional star allele-based haplotyping and GVB scoring methods for predicting the last cycle 6-mercaptopurine (6-MP) dose intensity percentage (DIP) as an indicator for 6-MP intolerance of ALL patients with *NUDT15* and/or *TPMT* deficiency. DIP prediction accuracies of GVB and star allele-based predictions were compared using AUROC (Area Under the Receiver Operating Curve) analysis. To define high-risk DIP groups, specificity, sensitivity, PPV, and NPV was computed under the binary classification model with nine different cutoff levels (*i.e.*, 5%, 10%, 15%, 25%, 35%, 45%, 60%, 80%, 100%). Furthermore, a comprehensive comparison of the

accuracy of GVB with the accuracies of the RVIS and GDI was performed in predicting the wide variety of functional gene subcategories using receiver operating characteristics (ROC) curve analysis. Comprehensive genomic characterizations of PGx, complex-disease, and Mendelian-disease genes were performed using the following seven molecular genetic features: number of paralogs, number of singletons, per-person mutability, PPI degree, CDS length, McDonald-Kreitman neutrality index (NI), and protein complexity. A condition-specific score adjustment scheme that could augment the performance by leveraging the genetic knowledge about underlying genetic architectures was suggested.

The 'computational' GVB exhibited as an improved or at least comparable predictor than the 'empirical' star allele-based haplotypes for determining subjects with increased risk of 6-MP intolerance in pediatric ALL patients measured by the last cycle 6-MP DIP (DIP $\leq$25 AUC_{GVB}=0.677, $AUC_{star-allele}$ = 0.645). The GVB score is considered to be a powerful gene-level scoring method for the prioritization of pharmacogenes, while the other gene-level scores performed best in prioritizing Mendelian-disease genes. A general outline of genetic condition-dependent analysis scheme, in which optimized strategies can

be developed by applying the condition-specific patterns of molecular genetic features, was exhibited. In the exploratory analysis, GVB can be used as an evaluation method which can aggregate the functional variants impact identified in novel candidate genes. The traditional two-gene model (*NUDT15* and *TPMT*) for predicting 6-MP DIP <25% was outperformed by the three-gene model that included *CRIM1*.

Overall, the GVB score—as a fully individualized and quantitative gene-level scoring system—can improve the ability to prioritize clinically important PGx variants and to understand the genetic architectures of common complex diseases. The findings of the present study suggest that different strategies are necessary depending on different genetic backgrounds for improving personal-genome interpretations in the context of pharmacogenetics and common- and rare-disease phenotypes in the era of personal genomics.

Keywords: Gene-level scores, Pharmacogenetics, Mendelian-disease, Complex-disease, Variant burden

Student number: 2014-21328